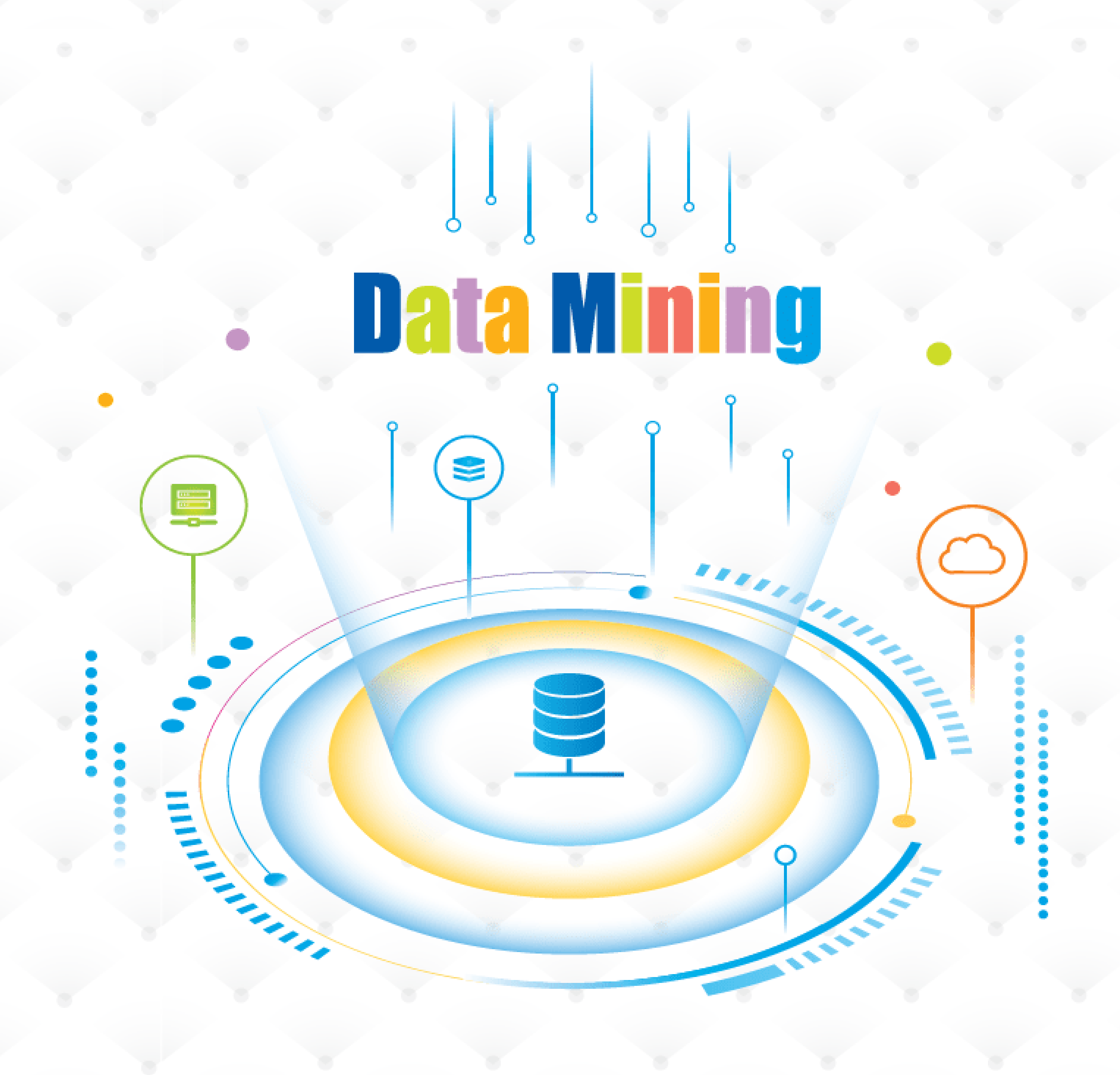


大数据技术与应用专业规划教材

数据挖掘实用案例分析

◎ 赵卫东 董亮 著



清华大学出版社

大数据技术与应用专业规划教材

数据挖掘实用案例分析

赵卫东　董　亮　著

清华大学出版社
北　京

内 容 简 介

数据挖掘已经广泛应用于各行各业，并催生了数据分析师的兴起。本书结合项目实践，首先对数据挖掘的核心问题进行了总结，并以保险推荐为例说明数据挖掘过程中每个步骤需要关注之处；然后，结合香水销售分析，讨论可视化图形的基本应用。为增强本书的实用性，提高读者的动手能力，后续章节详细地分析了数据挖掘在银行信用卡、餐饮、商务酒店、制造业、公安等领域的应用。此外，本书还介绍了卷积神经网络在音频数据处理方面的实际应用。

本书内容深入浅出，案例生动形象，可以作为高校相关专业“数据挖掘”“机器学习”“商务数据分析”等课程的实验教材，也可以供学习数据分析的社会人士参考。

图书在版编目（CIP）数据

数据挖掘实用案例分析/赵卫东，董亮著．—北京：清华大学出版社，2018
（大数据技术与应用专业规划教材）
ISBN 978-7-302-49049-4

Ⅰ．①数… Ⅱ．①赵… ②董… Ⅲ．①数据采集－案例 Ⅳ．①TP274

中国版本图书馆 CIP 数据核字（2017）第 295509 号

责任编辑：闫红梅 常建丽
封面设计：刘 键
责任校对：焦丽丽
责任印制：杨 艳

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 **邮 编**：100084
社 总 机：010-62770175 **邮 购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：北京国马印刷厂
经 销：全国新华书店
开 本：185mm×260mm **印 张**：16.5 **字 数**：400 千字
版 次：2018 年 2 月第 1 版 **印 次**：2018 年 2 月第 1 次印刷
印 数：1～1500
定 价：49.00 元

产品编号：075293-01

前言

FOREWORD

目前,高校的数据分析类课程(如数据挖掘、机器学习、大数据分析等)教学方式大多以“知识点”为核心组织教学,学生主要以学习知识为主,工程应用实践机会较少。教师将所要教授的知识点在课堂上讲述,课后再以作业练习、课程实验、课程设计等形式帮助学生深入理解课堂上所学的知识。尽管为提高教学效果,目前许多高校尝试了大型开放式网络课程(Massive Open Online Course,MOOC)、翻转课堂、移动课堂、同伴学习和小规模限制性在线课程(Small Private Online Course,SPOC)等教学方法的改革,但总体上来说,对于应用性较强的课程教学,还存在改进的空间,尤其是对学生的动手实践能力要求较高的数据分析类课程。现有的教学方法在传授理论知识时,缺少实际应用环节的支持,学生缺少在实际应用的背景下充分理解所学知识的机会,难以培养学生应用专业知识分析解决问题的技能和创新思维能力。

数据分析的方法是科学,但这些方法的选择和应用过程因问题而异,带有很强的艺术性。在现有专业课程教学模式下,学生仅仅了解需要学习基本的理论知识,缺少实践动手经历,难以获得这些知识的应用技巧,很少接触与企业实际项目相关的内容,因此学生的应用能力较弱,与企业实际的需求脱节。例如,在“数据分析”课程中,一般的教学方式是教师将具体数据分析的方法教授给学生,学生能够理解算法或方法的内容,但难以解决实际项目中应用具体算法碰到的问题。目前亟待克服数据分析类课程教学脱离企业所需能力的培养痛点,在课程学习的知识基础上,解决实际问题,引导学生解决数据分析实际问题的必要技能和思维方法。

实际上,数据分析绝大部分的教材和书籍还基本停留在基本理论和方法的介绍,实验部分的内容比较简单或者缺失,实际应用的内容不足。还有些实战性的书籍没有按照教材的方式编写,案例也比较粗略,数据分析过程中的一些技能解释肤浅。有关实际项目中数据分析过程思路的分析以及难点解析对教学,尤其是对实验或案例教学非常重要。最近几年,作者与多家企业合作,在数据分析领域辛苦耕耘,亲自参与了多个实际数据分析项目,熟悉数据分析过程的酸甜苦辣,希望通过本教材弥补国内数据分析实用教材的不足,也希望本教材的出版能改善国内数据分析类课程教学资料短缺的情况。

学习数据分析的最好方法就是做中学,使用实际数据解决实际问题,而不是单纯学习技术。实际上,有效的数据分析需要对业务进行深入理解,在此基础上形成有效的分析思路,并通过实验反复比较,才能真正解决客户的问题。在数据时代,现实应用中往往不乏数据。从生活中的小数据、简单问题开始,做各种假设,探索其中的规律。不断尝试常用的分析语言、工具和技术,在应用中不断学习新的知识,弥补课堂教学的不足,尤其是体会数据分析过

程中书本上难得看到的分析技巧，并在应用中举一反三。如此反复，随着分析问题的深入，不断提高分析能力，体会数据分析的艰辛和解决客户问题的快乐。

本教材不局限于数据分析基本理论和基本方法的介绍，而是立足实际应用，突出实际数据分析项目中的思路，以及数据分析中的难点。但希望读者具有一定的统计学、机器学习(数据挖掘)、数据科学，以及必要的相关专业知识。也不追求过多的案例堆积，希望读者能理解数据分析的思路，举一反三。这些内容是作者多年项目实践和教学成果的总结，其中的分析思路只有参与实际的项目，才能体验到数据分析的难点和艺术性，这是目前教学过程中培养学生工程性思维的重要问题，也是真正提高学生创新能力和动手能力的手段。这些内容是数据分析的基础，也是从事大数据分析必须掌握的知识和技能。有关数据挖掘常用算法的介绍，读者可以参阅作者已经出版的教材《商务智能(第 4 版)》(清华大学出版社，2016 年)或其他专业书籍。

全书分为 11 章，具体的内容简介如下：

第 1 章从数据分析的流程出发，讨论了在数据分析各个阶段需要做的工作以及经常遇到的主要问题，尤其是数据挖掘算法使用时容易遇到的难题。数据挖掘过程有一定的标准，但是针对具体的业务需求，如何设计合理、有效的数据分析流程，需要有一定的经验和技巧，数据的预处理、算法的选择等主要步骤都充分体现了数据挖掘的艺术性。

第 2 章以保险产品推荐项目为例，突出了数据挖掘选择合适的算法并非很简单的事情，需要在理解分析问题以及对多种算法熟悉的基础上，通过实验对初选的几种算法进行比较、调优，才能选择对解决问题效果比较好的算法。

第 3 章介绍了多维分析常用的可视化图形，这是数据分析的基本功。这些图形可以帮助数据分析师探索数据，找出数据中存在的问题以及基本规律。

第 4 章介绍了 IBM SPSS Modeler 18 数据挖掘工具的常用组件。在学习数据分析的不同阶段，根据学习者的基础、问题的分析难度等，可以选择不同的工具或平台。尽管分析工具并不是数据挖掘最重要的事情，但学习成本低、功能强大的分析工具对于问题的解决也是不可少的。对于编程基础有限的数据分析师，可以选择类似 IBM SPSS Modeler 18 的挖掘工具或 TensorFlow 等开源工具。尽管如此，对于有一定数据分析基础的读者，推荐学习 Python、R 等针对数据分析的语言，这些语言比较灵活，功能也十分强大。

第 5 章对香水的销售数据进行分析，讨论受欢迎的香水以及特点，并找出影响香水销售的主要因素，为香水的营销提供依据。

第 6 章对银行的客户信用记录、申请客户信息、拖欠历史记录、消费历史记录等人口属性、交易数据进行综合分析，讨论用户银行信用卡拖欠和欺诈行为特征，为银行推广信用卡以及风险管理提供依据。

第 7 章从大众点评网抓取火锅店海底捞的菜品介绍以及客户评论数据，以客户为中心，分析客户对火锅的偏好，为火锅店的选址、菜品的选择和设计，以及火锅店的竞争力都提供了参考。

第 8 章以携程网上某商务宾馆的客户评分、评论数据为基础，通过情感分析，分析了客户对商务宾馆的偏好，并了解客户的消费行为，比较多家商务宾馆的竞争优劣势，为商务宾馆改进经营提供了参考。

第 9 章在某耐热导线工厂最近 2 年的质量管理数据的基础上，分析了这些数据存在的

问题，探索耐热导线的加工流程中几个工序之间半成品或成品质量指标的关系，提高最终产品的合格率。

第10章利用公安人口数据和违法犯罪人员行为特点的数据，建立风险评分模型，实现对高危人群的特征分析，识别具有违法、犯罪、可疑或可能的高危人员。

第11章讨论深度学习在音频处理领域的应用，介绍了常用的深度神经网络模型，重点分析卷积神经网络在音频质量评价领域的应用。

数据挖掘是一个多学科交叉的领域，本书通过少数实际的具体案例，阐述数据分析项目的过程以及一些要点，可作为普通高等学校“数据挖掘”“商务数据分析”“商务智能”等课程的案例和实验指导材料，也可供有志于数据分析师的读者参考。配套实验数据、源代码、软件等可以从清华大学出版社网站下载。由于作者水平有限，书中难免有错误之处，希望读者不吝指出。

在写作的过程中，胡远文、于召鑫、黄黎明、蒲实、朱荣斌等在资料收集方面做了一些工作，在此表示感谢。

赵卫东
2017年8月
复旦大学

目录

CONTENTS

第 1 章

数据分析过程的主要问题

数据分析是一种入门容易但要精通却很难的学科。做好数据分析并非依赖于某一种技术或方法，其关键是分析思路，通过对业务进行调研，思考过程具有逻辑，并引入一定的创新理念，最后形成可行性建议。数据分析人员为了完成分析任务，获得较好的分析结果，不仅要懂得行业知识，对业务流程有一定的了解，还要理解数据背后的隐含信息，能够对数据进行合理的解读，而且要从变化的角度和时间维度对需求进行把握，确定用哪些数据来解决行业问题，这是数据分析的基础。

数据分析的主要流程是：明确分析目标、数据收集、数据预处理、建模分析、结果评估、结论整理及建议，通过对现状、原因等分析最终实现预测分析，确保数据分析维度的充分性和结论的合理有效性。

1.1 业务理解

数据分析过程中需要理解需求和分析目标，深入理解与分析目标相关联的业务背景，包括行业知识、领域知识及业务流程等，若数据分析人员对业务背景不熟悉，其分析方法和过程就难以贴合实际需求。业内专业人员往往以数据分析人员分析的结论为常识。

为了从数据中挖掘出有价值的结果，与领域专家进行充分交流，要亲临一线去了解业务实际情况，切忌“数据空想”，对业务知识理解其逻辑和原理，不仅有助于在数据预处理过程中对异常数据进行甄别和剔除，而且有助于分析过程中数据探索和挖掘方法的选择，对于结果是否符合预期，也可直观得出结论，否则容易出现模型的准确率虽然很高，经过业务专家评价时发现模型的某一自变量为目标变量的特征表现，最终模型毫无价值。

对数据分析目标的理解，包括定性分析和定量分析，前者给出与目标变量关联的自变量列表或目标变量的性质预测等，后者除了列举相关自变量，还要对其权重等进行定量分析，

在实际数据分析过程中,需要依据不同的业务目标设计分析方案。

在业务理解中,要以方法论的层面进行流程梳理,以实现快速确认分析目标相关联的影响因素,将分析过程以结构化的方式展现,利于理顺思路,而且不局限于某一行业应用,只要变换行业影响因素,即可应用于其他行业。例如,在企业经营活动的分析中,可以应用图 1.1 所示的分析框架,其中主要包括产业基础、企业运营分析、企业财务分析、竞争分析、营销分析、客户分析,此分析框架基本涵盖了大部分的企业经营活动,具体分析中可以适当进行增减和完善,并且可以按照不同的行业进行细化,形成行业分析框架。

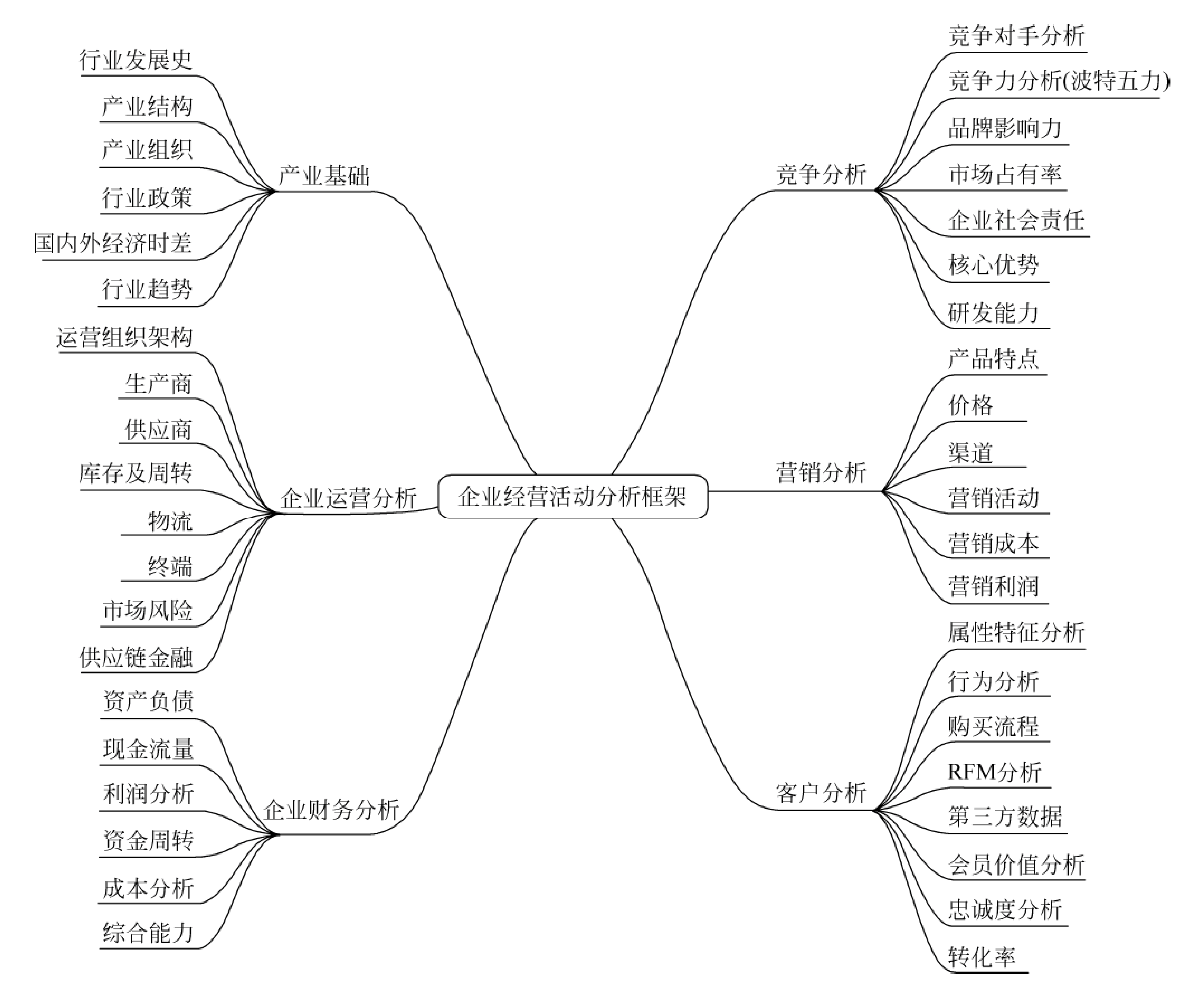


图 1.1　企业经营活动分析框架

对业务理解的分析框架中,主要从宏观的角度结构化、模块化指导数据分析,把问题分解成各个相关联的子模块,为后续数据分析进行规划,起到提纲挈领的作用。

1.2　数据理解

数据分析从字面上看是由数据和分析两部分组成的,其中数据是基础和根本,没有数据样本作为支撑,再好的结论也是无本之木,对现有数据理解到位有助于建立合理的分析框架。分析目标相关联的自变量数据往往可遇不可求,多数情况下,数据资料与分析的目标没

有直接相关性，需要对数据本身进行探索，查看其数据特性或样本特征，结合这些特征来挖掘其与分析目标之间的关系。

为了提高数据分析的准确性，需要多维的源数据，数据量较大可能会产生更多的冗余数据，处理过程较麻烦，经过预处理和降维后，可以得到更多样的支持数据，在初创型企业的数据量较少的情况下，可通过爬虫抓取非结构化数据，并转化为结构化数据作为补充。

了解业务流程中数据产生过程，明确数据代表的意义，并对数据的结构和各字段之间的关系进行分析，在分析过程中需要结合业务逻辑，对数据的理解是整个数据分析过程的基础，如果这一过程出现问题，将影响最终分析结果的正确性。

从历史的角度，数据的产生过程本身是变化的，在时间的维度上，不仅要关心数据是如何产生的及产生的频度，还要关心用户的动作数据，这些都将产生趋势特征，在数据分析过程中，需要关注业务变化导致的数据变化。

同时，由于需求会发生变化，新的数据会加入进来，数据分析方案也要具有一定的扩展性，以应对企业发展的变化和原始数据变化带来的影响，能够在设计模型后对其进行修正和动态改进。

1.3　数据质量问题与预处理

数据质量要求数据是完整的和真实的，并且具有一致性和可靠性。在数据分析过程中，高质量的数据更容易具有较高的区分度。相反，在数据分析领域，有一个著名的“垃圾进，垃圾出”结论，如果数据具有较多缺失值、异常值和无效记录，那么依此数据建立的模型在实际应用中将无法保证其结果真实和有效，数据预处理占用整个数据挖掘项目60%的工作量，目标就是保证输入模型的数据是符合业务实际情况的，基于正确的数据，才可以谈模型的选择和应用。

1. 数据量较少

数据挖掘需要有一定的数据量作为支撑，随着数据量的增多，其中的规律越发明显，也更容易发现其中分析目标相关的因素，特别是在神经网络或深度学习等算法中，其前提条件就要求有大量的训练数据，否则就容易引起模型过拟合的问题。

数据分析过程中一般要将样本划分为训练集、验证集、测试集，如果数据量较少，可以只需要训练集和测试集，其中训练集的数据量一般为50%～80%。在某些数据质量较高、区分度较明显的业务场景中，数据量可以更少，一般来说，数据量是自变量数量的10～20倍为佳。

在数据的数量足够多的情况下，还要关注数据的质量，如果给定的数据虽然较多，但其中样本的覆盖范围较少，与分析目标相关维度的数据数量才是关键的，否则最终分析得到的结论可能会有较大的局限，不能完全反映数据的本质。

2. 数据量过多

数据集中数量过多时，对全部数据集进行分析要耗费更多的计算资源，要求硬件配置较高，并且由于数据中各类数据的比例往往是不平衡的，例如，两家公司的产品销售的开始时间点并不一致，其销量相差悬殊，如果直接应用到模型中进行竞争分析，则可能出现较大的

结果误差，这种情况可以应用数据采样技术随机提取样本子集。

在面对海量的同质化数据时，如商品交易数据，可以通过聚集技术按照时间、空间等属性进行平均值等汇总，减少数据数量，由于采用了统计汇总后的数据，结果的可视化层次更高，也更加稳定，缺点是可能存在细节丢失的情况。

另外一种情况是在小概率事件的处理中需要关心数据集的不平衡问题。例如，在车辆运行异常检测时，车辆正常运行的时间远超过出现故障的时间，所以正常的数据量占了绝大多数，异常数据量极少，或者是在广告点击事件、地震检测、入侵检测、垃圾邮件过滤等这类稀有事件的分析中，要对数据集应用采样技术，或对异常数据进行复制，提高其占比。

3. 维度灾难

当数据中的自变量较多时，会出现维度灾难问题，特别是在矩阵数据中，其中冗余变量占比较高时，可用数据变成稀疏矩阵，在分类算法处理时就没办法可靠地进行类别划分，在聚类算法中则容易使聚类质量下降，为了从中获得稳定的分析结果，需要耗费大量的运算时间，分析过程低效，为了应对此问题，可以采用线性代数的相关方法将数据从高维空间影射到低维空间中，其中主成分分析(PCA)、奇异值分解(SVD)等方法比较常用。

下面通过对信用卡消费行为与是否存在欺诈进行分析，来展示 PCA 的主要用法。信用卡用户消费统计记录如图 1.2 所示，其中包括了卡类别、日均消费金额、日均次数等消费行为统计后的结果值，还包括用户的属性信息，如性别、年龄、职业等，排除目标字段，共有 19 个输入变量可供选择。

	卡类别	日均消费金额	日均次数	单笔消费最低	单笔消费最高	年收入	是否存在欺诈	性别	年龄	婚姻	户籍	教育	居住类型	职业	工作年限	保险缴纳	车辆情况	总评分	信用等级	额度
1	普卡	764	6	45.300	1127.200	54000	0	男	32	已婚	河南	本科	租房	个体户	9	有	无	60	D-风险客户	10000
2	普卡	797	2	48.000	1303.000	56000	0	女	34	已婚	河北	本科	租房	个体户	11	有	无	60	D-风险客户	10000
3	普卡	106	4	6.600	129.900	22297	0	女	60	已婚	北京	本科	自购房	私营企业	37	有	有	60	D-风险客户	10000
4	普卡	800	2	48.000	1308.000	56000	0	男	34	已婚	北京	本科	租房	私营企业	11	有	无	60	D-风险客户	10000
5	普卡	968	4	58.700	2411.900	67400	0	男	31	未婚	浙江	大专	租房	私营企业	11	有	无	60	D-风险客户	10000
6	普卡	101	3	5.000	102.000	19360	0	男	32	已婚	湖南	硕士及以上	自购房	私营企业	7	有	有	60	D-风险客户	10000
7	普卡	106	4	6.700	130.200	22599	0	女	45	已婚	上海	本科	自购房	私营企业	22	有	有	60	D-风险客户	10000
8	普卡	800	2	48.100	1315.300	56000	0	女	33	已婚	黑龙江	本科	租房	个体户	10	有	无	60	D-风险客户	10000
9	普卡	366	3	26.700	634.700	39198	0	女	34	已婚	海南	本科	租房	私营企业	11	有	无	60	D-风险客户	10000
10	普卡	107	2	6.900	131.500	22975	0	男	29	未婚	天津	硕士及以上	自购房	私营企业	4	有	有	60	D-风险客户	10000

图 1.2　信用卡用户消费统计记录

在 SPSS Modeler 中应用主成分分析/因子节点对数据进行降维，选择日均消费金额等 9 个字段作为输入，以 70%训练集、30%测试集的比例进行分区，选择“专家”模式，参数为默认值，运行后的主要结果如图 1.3 所示。

在总方差解释表中，前 4 个变量的初始特征值大于 1，分别为日均消费金额、日均次数、单笔消费最低、单笔消费最高，这 4 项累积占全部变量的 84.507%，也符合主成分的 80%以上的标准，说明这 4 项作为输入变量比较合理。

降低维度的另一种方法是通过特征子集选择的方式，将那些不相关的特征，如身份证号、姓名等剔除，只选择与目标变量紧密相关的特征。除了剔除属性，还可以使用特征加权技术，结合领域知识人为赋予某些特征更大的影响力权重。

在深度学习领域，常用特征提取和特征创建的技术将原始数据中的特征进行重构，以获得模型需要的特征，并且在重构过程中加以格式转换和数据变换。常用的技术包括傅里叶变换和小波变换，前者将时域信号转化为频域信号，后者主要处理时间序列等类型。

4. 数据不完整

除了数据量要多，还要求数据的种类要多。例如，要对企业产品的销售情况进行分析或预测，除了需要有企业产品相关的市场、销售情况等信息外，还需要有客户相关资料、竞品的

公因子方差

	初始	提取
日均消费金额	1.000	0.914
日均次数	1.000	0.697
单笔消费最低	1.000	0.922
单笔消费最高	1.000	0.785
年收入	1.000	0.553
年龄	1.000	0.927
工作年限	1.000	0.928
总评分	1.000	0.938
额度	1.000	0.942

提取方法：主成分分析法。

总方差解释

成分	初始特征值			提取载荷平方和		
	总计	方差百分比/%	累积/%	总计	方差百分比/%	累积/%
1	3.351	37.234	37.234	3.351	37.234	37.234
2	1.901	21.123	58.357	1.901	21.123	58.357
3	1.280	14.217	72.575	1.280	14.217	72.575
4	1.074	11.932	84.507	1.074	11.932	84.507
5	0.783	8.704	93.211			
6	0.304	3.379	96.590			
7	0.145	1.611	98.201			
8	0.112	1.243	99.444			
9	0.050	0.556	100.000			

提取方法：主成分分析法。

图 1.3　PCA 主成分分析结果示例

销售情况、市场数据、财务数据等，甚至要有交通物流、CPI 等宏观数据支持，但是现实情况中，很多数据缺失，要么这些数据并没有进行记录，要么它们在竞争对手的系统中，无法获得，这种情况将直接影响数据挖掘方法的选择，此时可以通过编写程序，来抓取外部数据作为补充。

数据缺失也是数据不完整的一种表现，可能是空白值或空值，也可能是存在大量的无效值，例如，所有记录的某一字段值均相同，或者某一字段中超过一半的记录为空或无效，在出现数据缺失时，分析人员要查找缺失原因，是原信息录入系统缺陷，还是人为操作失误，或者字段为选填等业务原因，并按照不同的原因进行数据预处理。例如，由于系统 Bug 导致的，则需要修复 Bug 并重新计算，如果当前字段中的数值是随时间逐渐生成的，则为业务原因，需要结合实际业务进行处理。

对缺失值可以采用众数、中位数、均值、最近距离等方法对缺失值进行人为补充，或者也可以通过回归或贝叶斯定理等预测缺失值。为了提高数据的纯度，也可以删除含有缺失值的记录，但如果缺失值的记录数较多时，删除操作可能会丢失样本特征，此时可以删除对应的字段，对于缺失值超过 30%的字段，可不作为模型输入变量。

5. 异常数据

在数据收集阶段由于人为或系统处理等原因，会导致产生异于常规的数据。其中异常数据分为两类：一类是错误的数据；另一类为小概率事件，或称为稀有事件。在系统预处

理阶段要视情况对数据进行探索，并结合行业内的业务知识对其进行识别，一旦发现错误数据，则将其剔除或修正。对于稀有事件，如信用卡欺诈行为、垃圾邮件等，这类正常数据不但不能修正和删除，反而要重点分析其特征。

通过查看散点图或箱图的方式查看离群点信息，如图 1.4 所示，可以看到方框中的年收入达到 21 亿元，已经超过绝大多数人的收入范围，极有可能为异常数据。还可基于距离或统计模型等进行检测，如应用线性回归、主成分分析等方式来区分异常数据，除此之外，还可应用深度学习(如 RNN 方法)来检测。

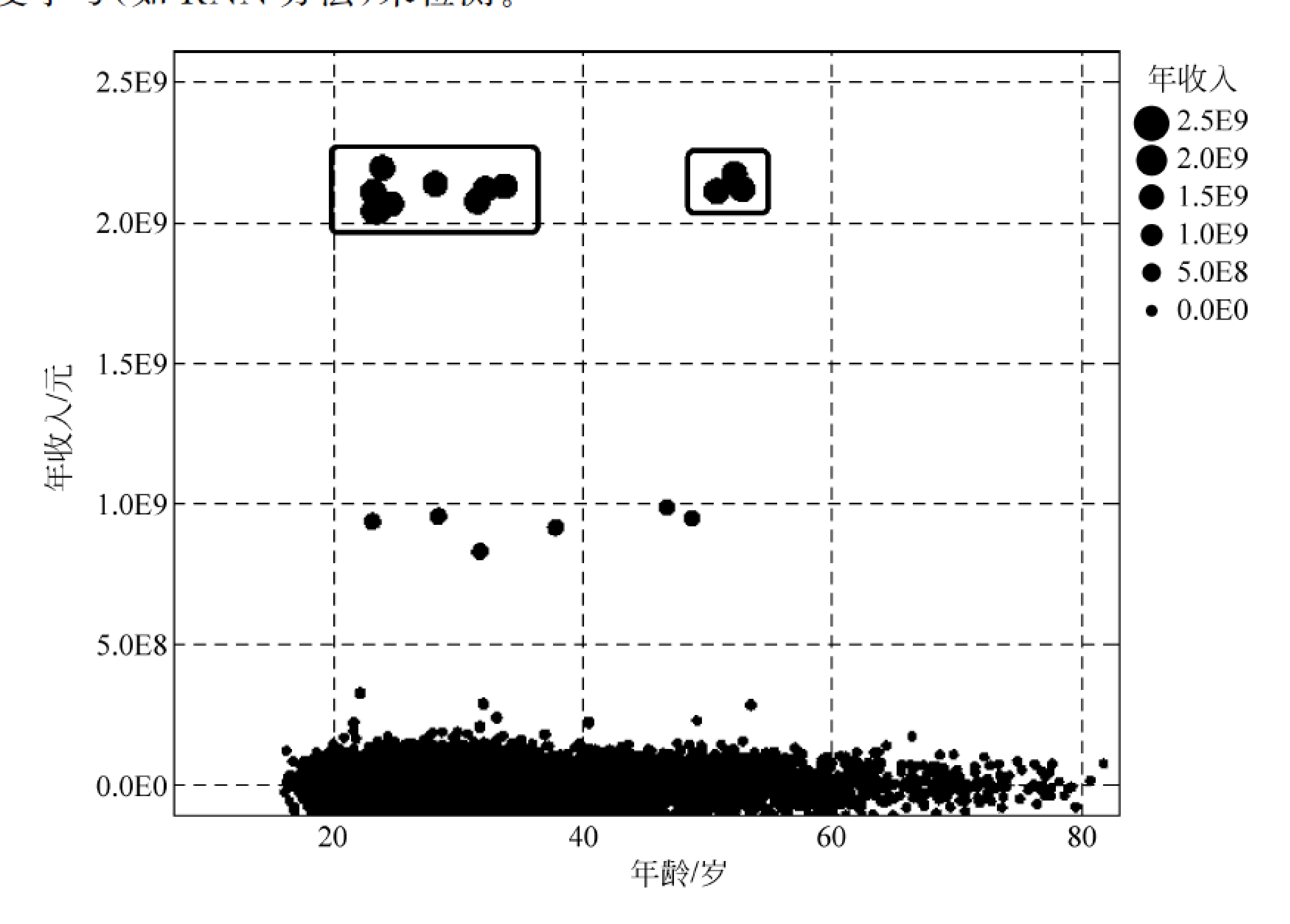


图 1.4 通过散点图查看离群点

当异常数据并非在离群点时，没有显著异常，可能是由于人为输入错误或系统误差导致的，虽然这些数值是不正确的，但是由于其与真实值之间区分较少，所以较难发现这类噪声数据。可以通过抽样的方式进行人工检测，或者对比不同数据源系统中的数据，进行一致性检测。

6. 重复数据

在数据分析中如果出现较多的重复数据，将对模型的结果产生误差，在数据处理过程中可以使用 SQL 或 Excel 中的去重复方法将重复数据滤除。有时候在记录中所有字段都是非重复数据，但选择其中部分字段时则容易产生重复样本，即样本子集中含有重复数据，特别是手动选取某几个字段作为模型输入时，容易忽略这一细节，所以，在将其应用到模型之前，需要进行过滤，将重复数据滤除。在 SPSS 中可以使用“区分”节点，对选择的自变量进行去重。利用“区分”节点去重复如图 1.5 所示。

在模式中选择“每组仅包括首个记录”，其他重复的记录将滤除，用于分组的字段即为流向下一节点的变量，只有日均消费金额等 4 个字段中的值均为重复时，才会被滤除。

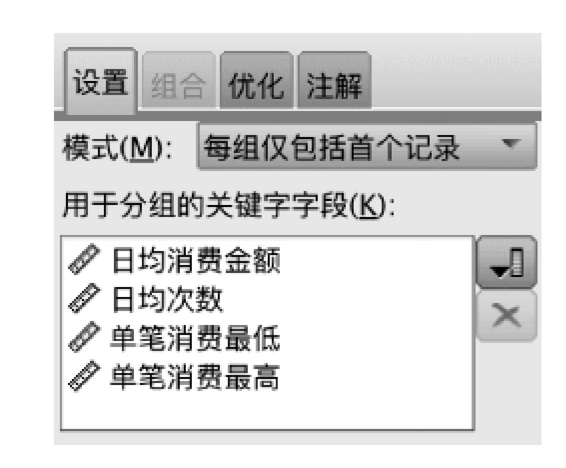


图 1.5 利用“区分”节点去重复

7. 数据不一致

随着数据源的增多，不同数据源中不同结构类型的数据可能会产生冲突，导致数据不一致或相互矛盾，也可能是由于名称或标识不同导致的，例如，中文和英文表示同一对象，或由于变量的统计口径不同导致。在数据处理中需要对其进行筛选，结合实际业务选择正确的数据，例如，对不同数据源的数据进行优先级标记，出现不一致的情况时优先使用某一数据源的样本。

数据不一致的另一个表现是记录中某些字段不符合规范，使其与数据逻辑之间存在不一致，可以按照数据使用规范建立合法性检测的规则，以此对数据进行验证。

对数据进行评估，主要是对数据的准确性、完整性、一致性等维度找出样本存在的问题，应用 SPSS 中的数据审核节点，可以查看相关异常情况。图 1.6 是数据审核节点的审核结果。从图 1.6 可以看到各自变量的类型及数据分布情况，以及极值、平均值、标准差、偏度、类别数（非连续型变量）、有效的记录数。其中，偏度用于对分布的不对称性进行度量。在变量中，数据呈正态分布时是对称的，所以其偏度值为 0。具有显著正偏度值的分布有很长的右尾。负偏度的分布有很长的左尾。当偏度值超过标准差的 2 倍时，则认为此变量不具有对称性。

审核　质量　注解

字段	样本图形	测量	最小值	最大值	平均值	标准 差	偏度	唯一	有效
年龄		连续	18	80	33.406	10.012	1.434	--	10000
工作年限		连续	0	50	12.420	10.178	1.305	--	10000
年收入		连续	10000	2147483647	4951478.864	77647845.675	25.788	--	10000
性别		标记	--	--	--	--	--	2	10000
婚姻		名义	--	--	--	--	--	4	10000
教育程度		名义	--	--	--	--	--	5	10000
职业		名义	--	--	--	--	--	5	10000
户籍		名义	--	--	--	--	--	30	10000
居住类型		名义	--	--	--	--	--	3	10000
车辆情况		标记	--	--	--	--	--	2	10000
保险缴纳		标记	--	--	--	--	--	2	10000
信贷情况		名义	--	--	--	--	--	6	10000
信用等级		名义	--	--	--	--	--	5	10000

图 1.6　数据审核节点的审核结果

在“质量”选项卡中查看数据质量，如图 1.7 所示，可以看到字段的完整性和完整记录比例，以及空值、字符型空值、空白、空白值、离群值、极值的数量等。

由于数据审核节点为输出节点，无法直接进行输出，但在 SPSS Modeler 中可以将数据审核节点生成数据准备节点向后传递，并设计数据验证规则、基准和偏度值等指标，其将对

审核 质量 注解

完整字段(%): 100%　完整记录(%): 100%

字段	测量	离群值	极值	操作	缺失插补	方法	完成百分比	有效记录	空值	字符型空值	空白	空白值
年龄	连续	166	0	无	从不	固定	100	10000	0	0	0	0
性别	标记	--	--	--	从不	固定	100	10000	0	0	0	0
婚姻	名义	--	--	--	从不	固定	100	10000	0	0	0	0
教育程度	名义	--	--	--	从不	固定	100	10000	0	0	0	0
职业	名义	--	--	--	从不	固定	100	10000	0	0	0	0
户籍	名义	--	--	--	从不	固定	100	10000	0	0	0	0
居住类型	名义	--	--	--	从不	固定	100	10000	0	0	0	0
车辆情况	标记	--	--	--	从不	固定	100	10000	0	0	0	0
保险缴纳	标记	--	--	--	从不	固定	100	10000	0	0	0	0
工作年限	连续	173	0	无	从不	固定	100	10000	0	0	0	0
年收入	连续	0	18	无	从不	固定	100	10000	0	0	0	0
信贷情况	名义	--	--	--	从不	固定	100	10000	0	0	0	0
信用等级	名义	--	--	--	从不	固定	100	10000	0	0	0	0
是否申请成功	名义	--	--	--	从不	固定	100	10000	0	0	0	0

图 1.7　数据审核节点中数据质量结果

异常数据进行标记，使得向后传递的数据保持较高的质量。

多重共线性是指多个自变量之间存在线性相关，当出现共线性问题时，模型的参数会变得不稳定，其预测结果的准确性大打折扣。多重共线性的检测分为视觉观察和定量分析两种方式，前者可使用交叉散点图来透视 N 维样本数据可能存在的问题，后者可使用回归分析的方法对共线性进行诊断。

输入变量为年龄、工作年限、年收入，为了方便查看各维之间的关系，在 SPSS 中使用图形板工具，选择散点图矩阵(SPLOM)作为可视化类型，运行之后可以直观看到工作年限与年龄呈现线性相关性，如图 1.8 所示。

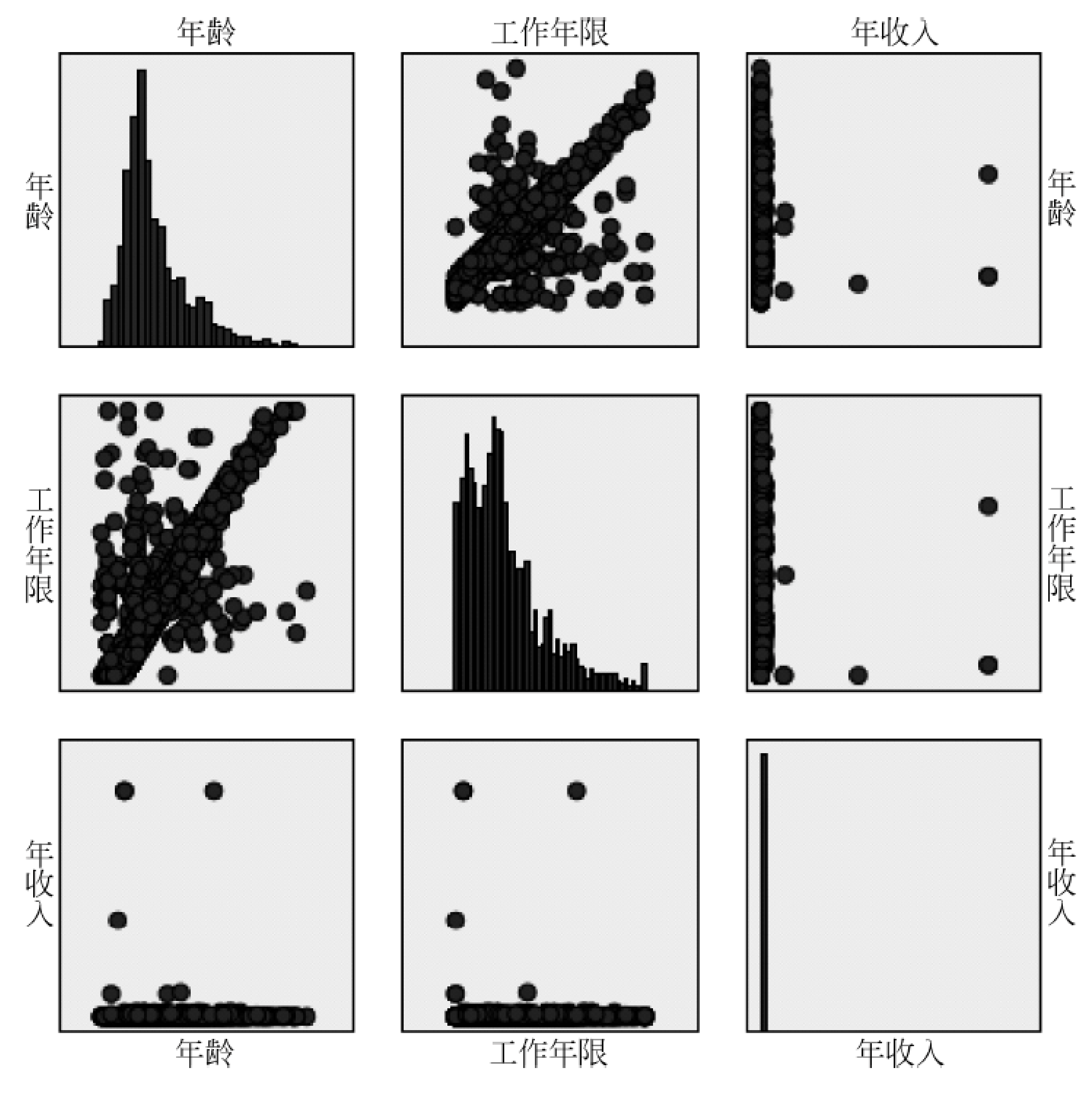


图 1.8　散点图矩阵透视数据

通过可视化的方式查找变量之间的共线性问题，只是初步的诊断，为了量化分析各变量之间是否存在较强的共线性问题，现在使用线性回归分析来检验，在回归分析的属性中以年收入为目标变量，以年龄、工作年限为输入，启用专家模式，并在输出中选中“共线性诊断”复选框，如图 1.9 所示。



图 1.9 回归分析检测共线性问题

运行回归分析之后，在模型结果的“高级”选项卡中查看共线性诊断结果，如图 1.10 所示，查看其条件指标发现第 3 维变量的条件指标达到 13.489(>10)了，说明年龄与工作年限确实存在线性相关。

共线性诊断

模型	维	特征值	条件指标	方差比例		
				(常量)	年龄	工作年限
1	1	2.752	1.000	0.01	0.00	0.01
	2	0.232	3.441	0.08	0.00	0.30
	3	0.015	13.489	0.91	1.00	0.68

图 1.10 回归分析诊断结果

在输入变量较多的情况下(本例中只有 2 个变量)，发现共线性变量后，将其剔除之后重新运行回归分析检验是否存在共线性问题，如果仍然存在，则继续排除字段，直到共线性问题不再出现。

1.4 数据分析常见陷阱

由于业务复杂度，数据多样，数据分析人员考虑不周等原因，在数据分析过程中会有很多陷阱，为了在应用中进行规避，这里列举几个常见的问题。

1. 错误理解相关关系

很多事物之间都存在相关性，但并不意味着其存在因果关系，或者有可能二者的因果关系颠倒了，要避免此类问题，一方面需要深入理解业务，规避大部分错误；另一方面要分析是否由第三方变量同时引起两种变量的变化，找出其变化原因。

2. 错误的比较对象

数据分析中的结果或效果比较时，容易将不同样本集进行结果比较，比较对象不合理，

其结果自然无效，结论便不能成立，这类问题很常见。例如，调查发现部队军人的死亡率要低于城市居民，但是分析人员没有对城市居民中的年龄等条件进行限制，二者并不具有同样的比较基础，所以其结论“参军很安全”自然也无法成立。

3. 数据抽样

在数据抽样时如果出现偏差，可能会影响分析结果，所以采样时，需要考虑什么时候进行采样，如何随机进行等，即按照什么标准来保证其子集能够代表全部样本，特别在分类问题中，目标类别的比例如果在采样时失去平衡，将直接影响分类结果。

4. 忽略或关注极值

有些时候，极值点或异常点是需要关注的，如果忽视它们，将可能失去某类样本或丢失某项重要特征，而如果在某些时候过于关注极值点，则可能会对结果造成偏差，影响结论。如何处理需要结合实际应用进行判断，要分析这些极值点出现的原因，从而决定其去留。

5. 相信巧合数据

有些数据分析结果会使人感到有一种假象，即结果恰好印证了之前的某个判断或猜想，实际上，如果重新进行多次实验，就会发现这不过是某种巧合而已。这类问题一般容易出现在医疗或生物学科领域中，或者是在回归分析中两个变量之间具有某种关联，可能是巧合。

6. 数据未作归一化

两个数据指标进行比较时，容易进行总数比较，而忽视比例的比较。例如，对比两个地区房价的增长情况，房屋单价同样涨 1000 元，上海可能涨幅只有 2%，而对于太原，可能达到 15%。忽视了总量对于指标的影响，必然影响结果的准确性。

7. 忽视第三方数据

我们在分析的时候往往只盯着手上的数据，由于维度有限，很多结论或观点是无法进行验证的，为了进一步深入分析，有必要搜集或使用爬虫获取更多种数据，使数据源更加丰富，这样也有利于比较分析，论证更加充分。

8. 过度关心统计指标

过于相信数据分析方法中的各项指标，就会忽视某些方法或结论成立的前提条件。例如，处理分类问题时，如果类别比例非常不平衡，99%为负例，只有 1%的正例，这种情况下，分类器一般不作分析，直接返回负例结果，准确率可以达到 99%，但是实际并没有意义，如果不加注意，可能会被指标欺骗。

1.5 数据分析方法的选择

数据分析方法要从业务的角度分析其目标，并对现有的数据进行探查，发现其中的规律，大胆假设并进行验证，依据各模型算法的特点选择合适的模型进行测试验证，分析并对比各模型的结果，最终选择合适的模型进行应用。

理解目标要求是分析方法选择的关键，首先对要解决的问题进行分类，如果数据集中有标签，则可进行监督式学习，反之可应用无监督学习方法。在监督式学习中对定性问题可用

分类算法，对定量分析可用回归方法，如逻辑回归或回归树等；在无监督式学习中，如果有样本细分，则可应用聚类算法，如需找出各数据项之间的内在联系，可应用关联分析。

熟悉各类分析方法的特性是分析方法选择的基础，不仅需要了解如何使用各类分析算法，还要了解其实现的原理，这样，在参数优化和模型改进时可减少无效的调整。在分析方法的选择过程中，由于分析目标的业务要求及数据支持程度差别较大，很难一开始就确认哪种分析方法效果最佳，需要对多种算法进行尝试和调优，尽可能提高准确性和区分度。

在选择模型之前，要对数据进行探索性分析，了解数据类型和数据特点，发现各自变量之间的关系，以及自变量与因变量的关系，特别注意在维度较多时容易出现变量的多重共现性问题，可应用箱图、直方图、散点图查找其中的规律性信息。

模型选择过程中先提出多个可能的模型，然后对其进行详细分析，并选择可用于分析的模型，在自变量选择时，大多数情况下需要结合业务手动选择自变量。选择模型后，比较不同模型的拟合程度，可统计显著性参数、R 方、调整 R 方、最小信息标准、BIC 和误差准则、Mallow's Cp 准则等。在单个模型中可将数据分为训练集和测试集，用来做交叉验证和分析结果的稳定性。反复调整参数，使模型结果趋于稳定。

1.5.1 分类算法

分类算法是应用规则对记录进行目标映射，将其划分到不同的分类中，构建具有泛化能力的算法模型，即构建映射规则来预测未知样本的类别。一般情况下，由于映射规则是基于经验的，所以其准确率一般不会达到 100%，只能获得一定概率的准确率，准确率与其结构、数据特征、样本的数量相关。

分类模型包括预测和描述两种。经过训练集学习的预测模型在遇到未知记录时，应用规则对其进行类别划分，而描述型的分类主要是对现有数据集中的特征进行解释并区分，其应用场景如对动植物的各项特征进行描述，并进行标记分类，由这些特征来决定其属于哪一类目。

主要的分类算法包括决策树、支持向量机（Support Vector Machine，SVM）、最近邻（K-Nearest Neighbors，KNN）、贝叶斯网络（Bayes Network）、神经网络等。

1. 决策树

正如其名，决策树是一种用于决策的树，目标类别作为叶子节点，特征属性的验证作为非叶子节点，而每个分支是特征属性的输出结果。决策过程是从根节点出发，测试不同的特征属性，按照结果的不同选择分支，最终转到某一叶子节点，获得分类结果。主要的决策树算法有 ID3、C4.5、C5.0、CART（可简写为 CART）、CHAID、SLIQ、SPRINT。图 1.11 是 C5.0 决策树算法应用实例，分析目标是信用卡申请是否成功的主要影响因素。

决策树的构建过程不需要业务领域的知识支撑，其构建过程是按照属性特征的优先级或重要性来逐渐确定树的层次结构，分支分裂的关键是要使其叶子节点尽可能“纯净”，尽可能属于同一类别，一般采用局部最优的贪心策略来构建决策树，即 Hunt 算法。决策树算法特点比较见表 1.1。

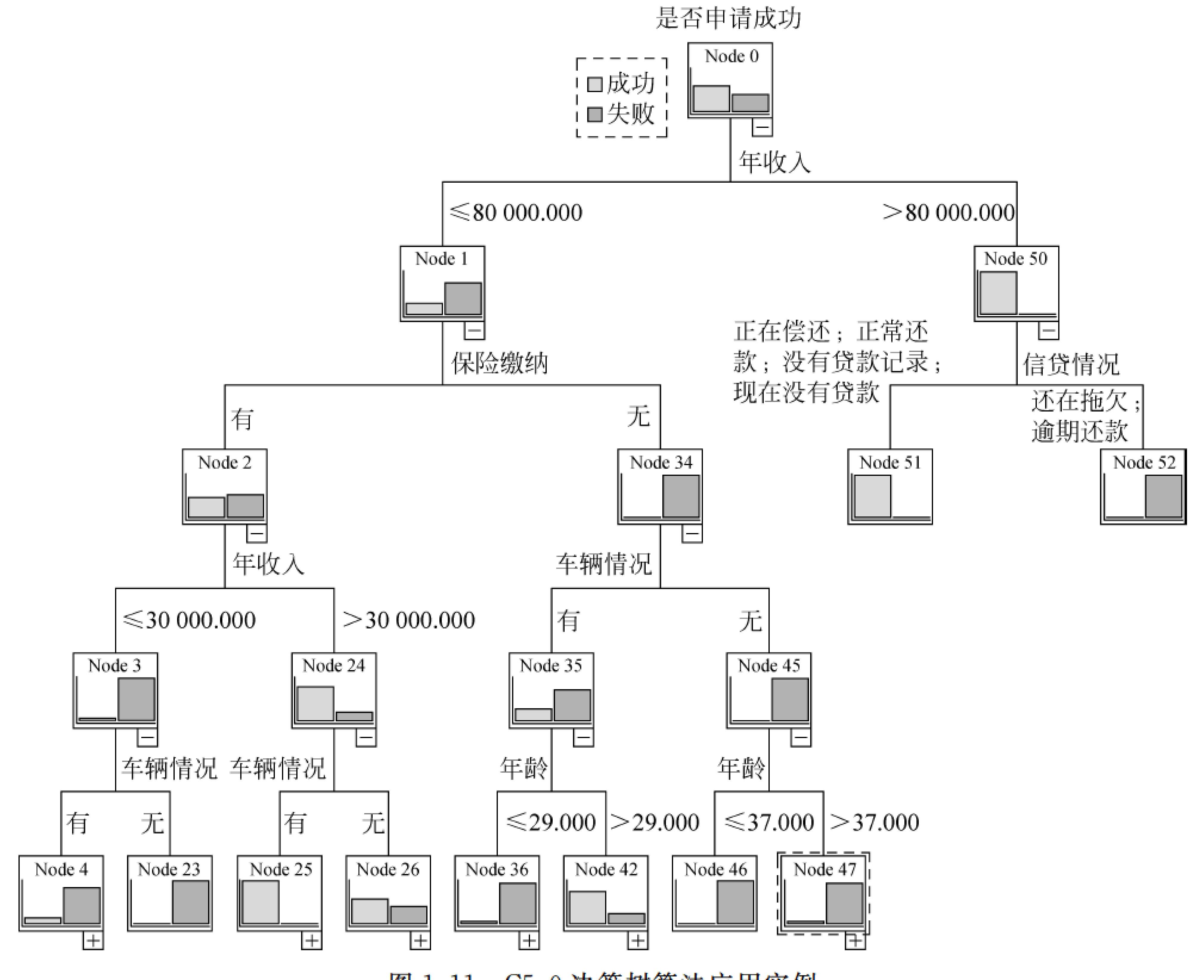


图 1.11　C5.0 决策树算法应用实例

表 1.1　决策树算法特点比较

决策树算法	特　　点		输出变量
ID3	优点	采用信息增益作为选择标准；整个决策树的熵值最小	分类
	缺点	只能处理离散变量；倾向于选择取值较多的属性；算法效率低	
C4.5	优点	采用信息增益率作标准；可处理不完整数据；规则易理解	分类
	缺点	数据集超过内存大小无法计算；多次扫描和排序，算法低效	
C5.0	优点	基于 C4.5 改进，更加稳健和准确，内存占用少；规则易于理解	分类
	缺点	输出变量必须为分类型	
CART	优点	自动忽略无贡献变量；训练时间短，且结果稳健；可以处理离散和连续属性	连续/分类
	缺点	对数值型输出变量的准确性低	
CHAID	优点	多分支树合并；按统计显著性确定分支变量和分割值	连续/分类
	缺点	无法处理大规模数据	
SLIQ	优点	采用广度优先构建树效率高；处理数据集较 C4.5 更大	分类
	缺点	数据集需常驻内存；算法复杂度与数据量呈非线性关系	
SPRINT	优点	减少常驻内存数据量；扫描效率高	分类
	缺点	难以对非分裂属性进行分裂；大数据集时需分批执行，效率低	
QUEST	优点	采用二元分类法，比 CART 更加简单、高效	分类
	缺点	目标字段须为分类；不能使用加权变量；有序字段须为数字型	
随机森林	优点	克服了过拟合；更加稳健；并行处理高维数据	连续/分类
	缺点	在某些噪声较大的分类或回归问题上会过拟合	

2. 支持向量机

支持向量机是由 Vapnik 等人设计的一种线性分类器准则，其主要思想是将低维特征空间中的线性不可分进行非线性映射转化为高维空间，使其线性可分。另外，应用结构风险最小理论在特征空间最优分割超平面，可以找到尽可能宽的分类边界，特别适合两个分类不容易分开的情况。例如，在二维平面图中某些点是杂乱排列的，无法用一条直线分为两类，但是在三维空间中，通过一个平面可以将其完美划分。

为了避免在低维空间向高维空间转化过程中增加计算复杂性和"维数灾难"，SVM 通过应用核函数的展开原理，不需要关心非线性映射的显式表达式，直接在高维空间建立线性分类器，极大优化了计算复杂度。SVM 常见的核心函数有 4 种，分别是线性核函数、多项式核函数、径向基函数、二层神经网络核函数。

SVM 的目标变量以二分类最佳，虽然可以用于多分类，但效果不好。相较于其他分类算法，在小样本数据集中其效果更好。由其原理可知，SVM 擅长处理线性不可分的数据，并且在处理高维数据集时具有优势。

3. 最近邻

通过在样本实例之间应用向量空间模型，将相似度高的样本分为一类，应用训练得到的模型对新样本计算与之距离最近(最相似)的 k 个样本的类别，那么新样本就属于 k 个样本中的类别最多的那一类。可以看出，影响分类结果的 3 个因素分别为距离计算方法、最近的样本数量 k 值、距离范围。

KNN 支持多种相似度距离计算方法：欧式距离(Euclidean Distance)、曼哈顿距离(Manhattan Distance)、切比雪夫距离、闵可夫斯基距离(Minkowski Distance)、标准化欧氏距离(Standardized Euclidean Distance)、马氏距离(Mahalanobis Distance)、巴氏距离(Bhattacharyya Distance)、汉明距离(Hamming Distance)、夹角余弦(Cosine)、杰卡德相似系数(Jaccard Similarity Coefficient)、皮尔逊相关系数(Pearson Correlation Coefficient)。

在 k 值选择中，如果设置较小的 k 值，说明在较小的范围中进行训练和统计，误差较大且容易产生过拟合的情况；k 值较大时意味着在较大的范围中学习，可以减少学习的误差，但是在其统计范围变大了，说明模型变得简单了，容易在预测的时候发生分类错误。

KNN 算法的主要缺点是：在各分类样本数量不平衡时误差较大；由于其每次比较要遍历整个训练样本集来计算相似度，所以分类的效率较低，时间和空间复杂度较高；k 值的选择不合理，可能会导致结果的误差较大；在原始 KNN 模型中没有权重的概念，所有特征采用相同的权重参数，这样计算出来的相似度易产生误差。

4. 贝叶斯网络

贝叶斯(Bayesian)网络又称为置信网络(Belief Network)，是基于贝叶斯方法绘制的、具有概率分布的有向弧段图形化网络，其理论基础是贝叶斯公式，网络中的每个点表示变量，有向弧段表示两者间的概率关系。

相较于神经网络，网络中的节点都具有实际的含义，节点之间的关系比较明确，可以从贝叶斯网络中直观看到各变量之间的条件独立和依赖关系，可以进行结果和原因的双向推理。图 1.12 是贝叶斯网络的一个示例网络结构，分析信用卡成功申请的影响因素。其中，"是否申请成功→年收入"表示：P(年收入|是否申请成功＝Yes)，即在申请成功的情况下，

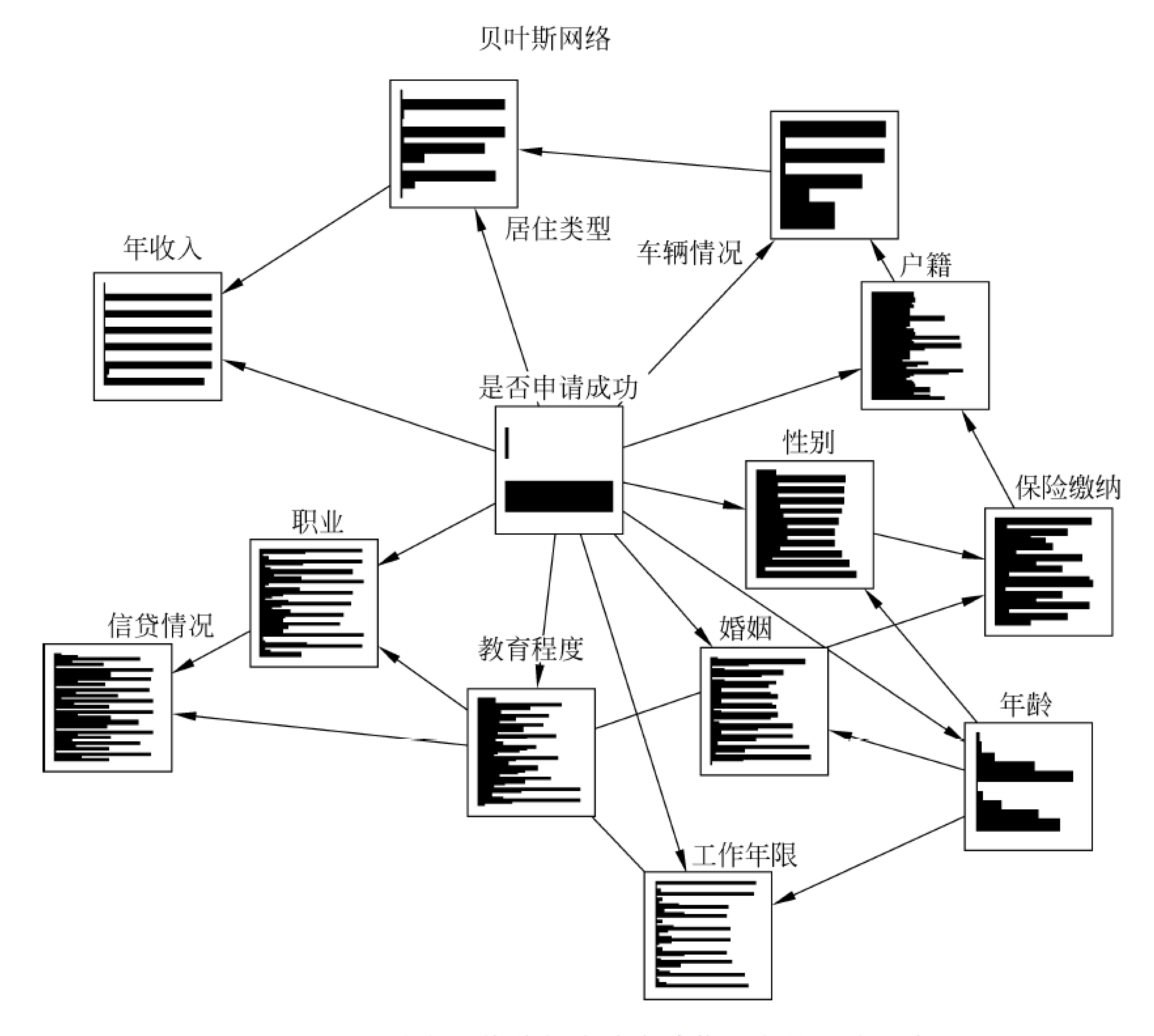


图 1.12　贝叶斯网络分析成功申请信用卡的影响因素

客户年收入的概率。

贝叶斯网络分类算法分为简单(朴素)贝叶斯算法和精确贝叶斯算法。在节点数较少的网络结构中,可选精确贝叶斯算法,以提高精确概率。在节点数较多的网络结构中,为减少推理过程和降低复杂性,一般选择简单贝叶斯算法。

5. 神经网络

传统的神经网络为 BP(Back Propagation)神经网络,目前的递归神经网络(RNN)、卷积神经网络(CNN)等均为神经网络在深度学习方面的变种,其基础还是由多层感知器(MLP)的神经元构成,这里仅介绍 BP 神经网络的特点,基本的网络中包括输入层、隐藏层、输出层,每一个节点代表一个神经元,节点之间的连线代表了权重值,输入变量经过神经元时会运行激活函数,对输入值按照权重和偏置进行计算,将输出结果传递到下一层中的神经元,而权重值和偏置是在神经网络训练过程中不断修正得到的。

神经网络的训练过程主要包括前向传输和逆向反馈,前者是将输入变量逐层向下传递,最后得到一个输出结果,并对比实际的结果,如果发现预测结果与实际结果不符,则逐层逆向反馈,对神经元中的权重值和偏置进行修正,然后重新进行前向传递结果,以此反复迭代,直到最终预测结果与实际结果一致。

BP 神经网络的结果准确性与训练集的样本数量和分类质量有关,如果样本数量过少,可能会出现过拟合的问题,无法泛化新样本;对训练集中的异常点比较敏感,需要分析人员对数据做好预处理,如数据标准化、去除重复数据、移除异常数据,从而提高 BP 神经网络的

性能。

由于模型结果神经网络是基于历史的数据构建的分析模型，如果是新数据产生的新规则，则可能出现不稳定的情况，需要进行动态优化。例如，随着时间变化，应用新的数据对模型进行重新训练，来调整网络的结构和参数值。

1.5.2 聚类算法

聚类是基于无监督学习的分析模型，不需要对原始数据进行标记，按照数据的内在结构特征进行聚集形成簇群，从而实现数据的分离，其中聚集的方法就是记录之间的区分规则。聚类与分类的主要区别是其并不关心数据是什么类别，而是把相似结构的数据聚集起来形成某一类簇。

在聚类的过程中，首先选择有效特征存于向量中，必要时将特征进行提取和转换，获得更加突出的特征，然后按照欧式距离或其他距离函数进行相似度计算，并划分聚类，通过对聚类结果进行评估，逐渐迭代生成新的聚类。

聚类应用领域广泛，可用于企业发现不同的客户群体特征、消费者行为分析、市场细分、交易数据分析等，也可用于生物学的动植物种群分类、医疗领域的疾病诊断、环境质量检测等，还可用于互联网和电商领域的客户分析、行为特征分类等。在数据分析过程中可以先用聚类对数据进行探索，发现其中蕴含的类别特征，然后再用其他方法对样本进一步分析。

按照聚类方法分类，可分为基于层次的聚类(Hierarchical Method)、基于划分的聚类(PArtitioning Method，PAM)、基于密度的聚类、基于机器学习的聚类、基于约束的聚类、基于网络的聚类等。

基于层次的聚类是将数据集分为不同的层次，并将其按照分解或合并的操作方式进行聚类，主要包括 BIRCH(Balanced Iterative Reducing and Clustering Using Hierarchies)、CURE(Clustering Using Representatives)等。

基于划分的聚类是将数据集划分为 k 个簇，并对其中的样本计算距离，以获得簇中心点，然后以簇的中心点重新迭代计算新的中心点，直到 k 个簇的中心点收敛为止。基于划分的聚类包括 K 均值(K-Means)等。

基于密度的聚类是根据样本的密度不断增长聚类，最终形成一组“密度连接”的点集，其核心思想是，只要聚类簇之间的密度低于阈值，就将其合并成一个簇，它可以过滤噪声，聚类结果可以是任何形状，不必为球形，主要包括 DBSCAN(Density-Based Spatial Clustering of Application with Noise)、OPTICS(Ordering Points To Identify the Clustering Structure)等。

1. BIRCH 算法

BIRCH 算法是指利用层次方法来平衡迭代规则和聚类，它只需要扫描数据集一次，便可实现聚类，它利用了类似 B+树的结构对样本集进行划分，叶子节点之间用双向链表进行链接，逐渐对树的结构进行优化获得聚类。

其主要优点是空间复杂度低，内存占用少，效率较高，能够对噪声点进行滤除，缺点是其树中节点的聚类特征树有个数限制，可能会产生与实际类别个数不一致的情况；对样本有一定的限制，要求数据集的样本是超球体，否则聚类的效果不佳。

2. CURE 算法

传统的基于层次聚类的方法得到的是球形的聚类，对异常数据较敏感，而 CURE 算法是使用多个代表点来替换层次聚类中的单个点，算法更加健壮，并且在处理大数据时采用分区和随机取样，使其处理大数据量的样本集时效率更高，且不会降低聚类质量。

3. K 均值算法

传统的 K-Means 算法的聚类过程是在样本集中随机选择 k 个聚类质心点，对每个样本计算其应属于的类，在得到类簇之后重新计算类簇的质心，循环迭代，直到质心不变或收敛。K-Means 存在较多变体和改进算法，如初始化优化 K-Means＋＋算法、距离优化 Elkan K-Means 算法、K-Prototype 算法等。

K-Means 算法的主要优点是：可以简单快速处理大数据集，并且是可伸缩的，当数据集中结果聚类之间是密集且区分明显时，聚类效果最好。缺点是：必须先给定 k 值，即聚类的数目，大部分时间分析人员并不知道应该设置多少个聚类。另外，K-Means 算法对 k 值较敏感，如果 k 值不合理，可能会导致结果局部最优（不能保证全局最优）。

4. DBSCAN 算法

DBSCAN 算法的目标是：过滤低密度区域，发现稠密度样本点。与传统的基于层次的聚类和划分聚类的凸形聚类簇不同，其输出的聚类结果可以是任意形状的聚类。主要优点是：与传统的 K-Means 相比，是不需要输入要划分的聚类个数；聚类结果的形状没有偏倚；支持输入过滤噪声的参数。

DBSCAN 的主要缺点是：当数据量增大时，会产生较大的空间复杂度；当空间聚类的密度不均匀、聚类间距差很大时，聚类质量较差。

5. OPTICS 算法

在 DBSCAN 算法中，初始参数 E（邻域半径）和 minPts（E 邻域最小点数）需要用户手动设置，这两个参数较关键，不同的取值将产生不同的结果。而 OPTICS 克服了上述问题，为聚类分析生成一个增广的簇排序，代表了各样本点基于密度的聚类结构。

1.5.3 关联分析

关联分析（Associative Analysis）通过对数据集中某些属性同时出现的规律和模式来发现其中的属性之间的关联、相关、因果等关系，其典型的应用是购物篮分析，通过分析购物篮中不同商品之间的关联，分析消费者的购买行为习惯，从而制定相应的营销策略，为商品促销、产品定价、位置摆放等提供支持，并且可用于不同消费者群体的划分。关联分析主要包括 Apriori 算法和 FP-growth 算法。

1. Apriori 算法

Apriori 算法的主要实现过程是：首先生成所有频繁项集，然后由频繁项集构造出满足最小信任度的规则。Apriori 算法依赖的重要性质是频繁项集的非空子集也是频繁项集。

由于 Apriori 算法要多次扫描样本集，需要由候选频繁项集生成频繁项集，在处理大数据量数据时效率较低，其只能处理分类变量，无法处理数值型变量。

2. FP-growth 算法

为了改进 Apriori 算法，Jiawei Han 等人提出基于 FP 树生成频繁项集的 FP-growth 算法，该算法只进行两次数据集扫描，且不使用候选项集，直接按照支持度构造出一个频繁模式树，用这棵树生成关联规则，在处理大数据集时效率比用 Apriori 算法大约快一个数量级，对于海量数据，可以通过数据划分、样本采样等方法进行再次改进和优化。

1.5.4　回归分析

回归分析是一种研究自变量和因变量之间关系的预测模型，用于分析当自变量发生变化时，因变量的变化值。要求自变量不能为随机变量，需要具有一定的相关性。可以将回归分析用于定性预测分析，也可以用于定量分析各变量之间的相关关系。

1. 线性回归

应用线性回归进行分析时，要求自变量是连续型或离散型的，因变量则为连续型的，线性回归用最适直线（回归线）去建立因变量 Y 和一个或多个自变量 X 之间的关系。

其主要的特点是：

(1) 自变量与因变量之间必须有线性关系。

(2) 多重共线性、自相关和异方差对多元线性回归的影响很大。

(3) 线性回归对异常值非常敏感，其能严重影响回归线，最终影响预测值。

(4) 在多元的自变量中，可以通过前进法、后退法和逐步法去选择最显著的自变量。

图 1.13 是上海私家车车牌拍卖中，竞拍警示价和最后平均成交价之间的关系。由于参与沪牌额度拍卖的人数较多，警示价人为干预了最终的竞拍价格，所以呈现出极强的线性关系。

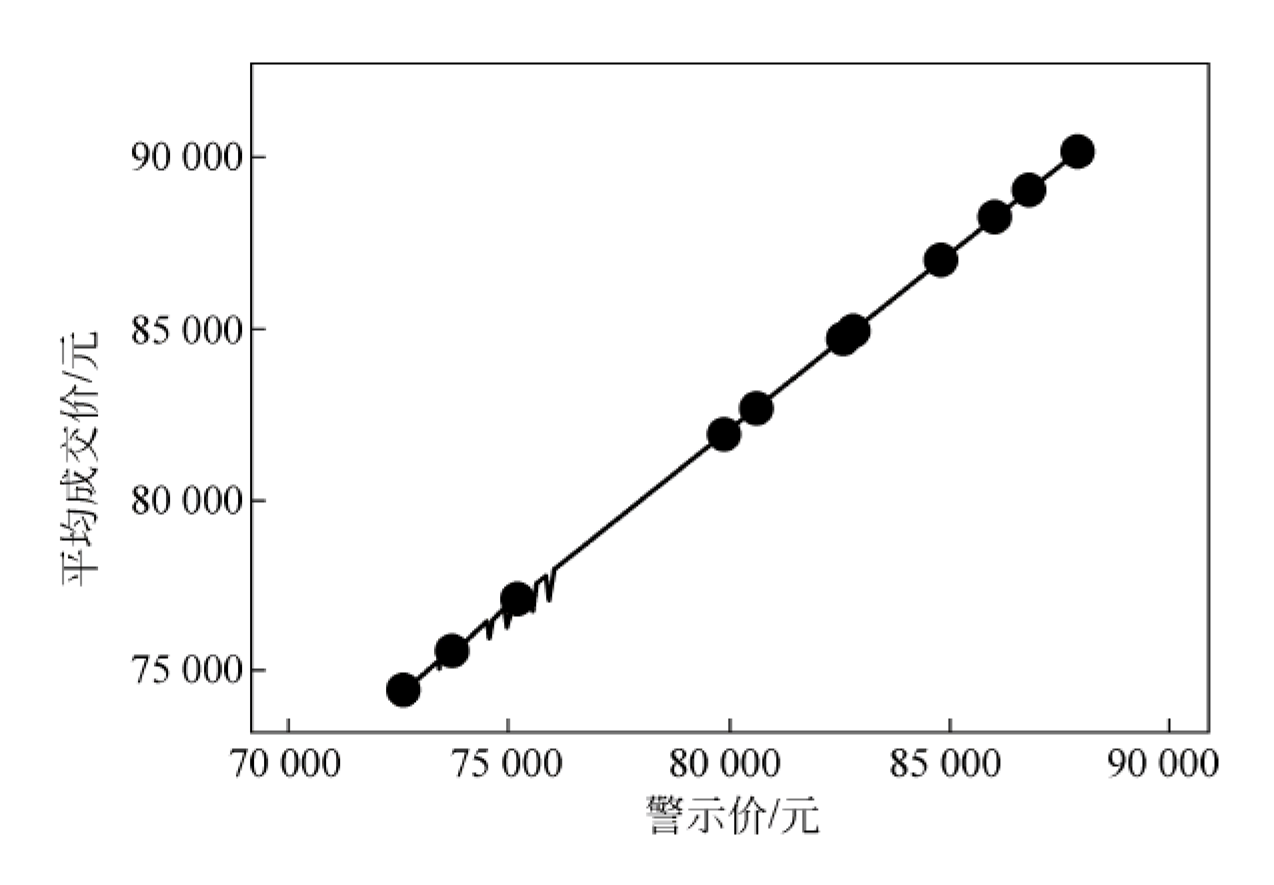


图 1.13　沪牌拍卖警示价与平均成交价呈线性关系

虽然其线性回归的结果准确率达到 98% 以上，但是实际应用中依然无法准确预测最终的平均成交价格，原因是成交价格的预测误差范围要求在 300 元以内，说明在实际分析中要与目标问题的环境要求相符合，而不是一味追求高准确率。

2. 逻辑回归

逻辑回归一般应用在分类问题中，如果因变量类型为序数型的，则称为序数型逻辑回

归，如果因变量为多个，则称为多项逻辑回归。逻辑回归的主要特点是：

相较于线性回归，逻辑回归应用非线性对数转换，使自变量与因变量之间不一定具有线性关系才可以分析。

为防止模型过拟合，要求自变量是显著的，且自变量之间不能存在共线性。可以使用逐步回归法筛选出显著性变量，然后再应用到逻辑回归模型中。

逻辑回归需要大样本量，在低样本量的情况下效果不佳，因为最大似然估计在低样本数量时其统计结果误差较大。

3. 多项式回归

在回归分析中，有时会遇到线性回归的直线拟合效果不佳，如果发现散点图中数据点呈曲线状态显示时，可以考虑使用多项式回归来分析。使用多项式回归可以降低模型的误差，从理论上多项式可以完全拟合曲线，但是如果处理不当，易造成模型结果过拟合，在分析完成之后需要对结果进行分析，并将结果可视化，以查看其拟合程度。

4. 逐步回归

处理多个自变量时，需要用逐步回归的方法自动选择显著性变量，不需要人工干预，其思想是：将自变量逐个引入模型中，并进行 F 检验、t 检验等来筛选变量，当新引入的变量对模型结果没有改进时，将其剔除，直到模型结果稳定。

逐步回归的目的是保证所有自变量集为最优的。用最少的变量去最大化模型的预测能力，它也是一种降维技术。主要的方法有前进法和后退法，前者以最显著的变量开始，逐渐增加次显著变量。后者是逐渐剔除不显著的变量。

5. 岭回归

岭回归又称为脊回归，在共线性数据分析中应用较多，是一种有偏估计的回归方法，在最小二乘估计法的基础上做了改进，通过舍弃最小二乘法的无偏性，以损失部分信息为代价使得回归系数更稳定和可靠。其 R 方值会稍低于普通回归分析方法，但其回归系数更加显著，主要用于变量之间存在共线性和数据点较少时的情况。

6. LASSO 回归

LASSO 回归的特点与岭回归类似，在拟合模型的同时进行变量筛选和复杂度调整。变量筛选是逐渐把变量放入模型，从而得到更好的自变量组合。复杂度调整是通过参数调整来控制模型的复杂度，如减少自变量数量等，从而避免过度拟合。

LASSO 回归也是擅长处理多重共线性或存在一定噪声和冗余的数据。可以支持连续型因变量，二元、多元离散变量的分析。

7. ElasticNet 回归

ElasticNet 回归结合了 LASSO 回归和岭回归的优点，同时训练 L1 和 L2 作为惩罚项在目标函数中对系统约束进行约束，所以其模型的表示系数既有稀疏性，又有正则化约束，特别适用于许多自变量是相关的情况，这时，LASSO 回归会随机选择其中一个变量，而 ElasticNet 回归则会选择两个变量。相较于 LASSO 回归和岭回归，ElasticNet 回归更稳定，且在选择自变量的数量上没有限制。

1.5.5　深度学习

深度学习方法是通过构建多个隐藏层和大量数据来学习特征，从而提升分类或预测的准确性，与传统的神经网络相比，不仅在层数上更多，而且采用了逐层训练的机制来训练整个网络，以防出现梯度扩散。深度学习包括了卷积神经网络（CNN）、深度神经网络（DNN）、循环神经网络（RNN）、对抗神经网络（GAN）以及各种变种网络结构。其本质是对训练集数据进行模式识别及特征提取和选择，然后应用于样本的分类。

目前，深度学习的方法在图像和音视频的识别、分类和模式检测等领域已经非常成熟，除此之外，还可以用于衍生成新的训练数据，以构建对抗网络（GAN），从而利用两个模型之间互相对抗提高模型的性能。

在数据量较多时可考虑采用这一算法，应用深度学习的方法进行分析时，需注意训练集（用于训练模型）、开发集（用于在开发过程中调参和验证）、测试集的样本比例，一般以6∶2∶2的比例进行分配。另外，采用深度学习进行分析时对数据量有一定的要求，如果数据量只有几千或几百条，极易出现过拟合的情况，其效果不如使用SVM等分类算法。

常见的权重更新方式为SGD和Momentum。参数初始值设置不当容易引起梯度消失或梯度爆炸问题；随着训练时间的推移，可以逐渐减少学习率。

1.5.6　统计方法

统计方法是在基于传统的统计学、概率学知识对样本集数据进行统计分类，是数据分析的基本方法，如对基于性别的数据进行分类、对年龄分段统计等。统计方法虽然看起来比较简单，但是在数据探索阶段尤其重要，可以发现一些基本的数据特征。分析技术并没有高深简易之分，与业务相结合、实用方便才是关键，所以不要小看传统的统计方法。经过认真细致的分析探索，一样可以发现数据中蕴藏的有价值的规律。

统计方法源于用小样本集来获得整体值集的各种特征，主要的统计方法或指标包括频率度量（如众数指标）、位置度量（如均值或中位数）、散度度量（如极差、方差、标准差等）、数据分布情况度量（如频率表和直方图）、多元汇总统计（如相关矩阵和协方差矩阵）。

根据汇总统计中置信度的计算方法，置信度达到95％以上，误差为－2.5％～2.5％，即置信区间宽度为5％，在汇总统计中需要的样本数至少为1000个，样本数量越多，其误差越小，所以在此类分析中要尽可能多地收集数据。

在描述统计分析时，往往会对不同维度进行样本分拆，划分越细，样本的纯度越高，信息就更有效，所以其结论的准确率就会越高，但是需要注意，分拆之后子维度的样本数量不能过少，否则结论过低会失去统计意义。

1.6　数据分析结果的评价

分析算法及其衍生的算法有很多，不同的算法具有不同的特点，并且在不同的数据集上表现也不一样，所以对分析结果的评价很重要，这样才能够知道在何种情况下选择何种算法，使用何种标准能达到分析的目标。

对结果进行分析时,常见的问题是容易混淆因果关系和相关性,例如,我们分析发现保养比较规律的汽车比保养维修不规律的出现意外事故的概率低,我们就认为保养规律与不发生意外事故呈现因果关系,而实际上可能是因为保养规律的驾驶人更自律,或者是其更加认真遵守交通规则,与是否发生意外事故只是相关关系。

在模型评价中容易出现主观性问题,由于数据采集或业务理解的局限,容易让分析人员认为某种方案的改进一定可以解决企业的问题,没有综合数据、业务、场景等多个维度对模型分析结果进行解读。分析报告虽然很有逻辑性,看起来很合理,但是不符合企业实际应用场景,反而对企业决策产生负面作用。所以,分析结果的评估需要业务专家参与,对结果的合理性、理解性、实用性进行评估,使其具有落地的价值。

1.6.1 分类算法的评价

对分类算法的结果评价主要有精确率(Precision)、F-Score、准确率(Accuracy)、召回率(Recall)、特效度(Specificity)、ROC(Receiver Operating Characteristic)曲线、曲线包围面积(Area Under Curve,AUC)。

上述指标涉及混淆矩阵的概念,如图 1.14 所示,其中总记录数 Total 为 4217 条,其中 TP 为 13 条,FP 为 175 条,FN 为 3 条,TN 为 4026 条,其中,精确率(Precision)是模型精确性的度量,预测正例数占所有正例数的比例,Precision=TP/(TP+FP)=13/(13+175)=0.07,准确率(Accuracy)是所有预测正确的记录数与总记录数之比,Accuracy=(TP+TN)/Total=(13+4026)/(13+3+175+4026)=0.96,召回率(Recall)是模型覆盖面的度量,是表示多少个正例被识别为正例,体现了分类器对正类的识别能力,本例中,Recall=TP/(TP+FN)=13/(13+3)=0.81,特效度(Specificity)是表示所有负例被识别正确的比例,度量的是对负例的识别能力,Specificity=TN/(FP+TN)=4026/(4026+175)=0.96。

实测	预测	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

混淆矩阵

实测	预测		
	Yes	No	比例正确
Yes	TP=13	FP=175	0.07
No	FN=3	TN=4026	1.00
比例正确	0.81	0.96	0.96

图 1.14 混淆矩阵示例

图 1.14 中,TP(True Positive)表示样本的真实类别为正,最后预测得到的结果也为正。FP(False Positive)表示样本的真实类别为负,最后预测得到的结果却为正。FN(False Negative)表示样本的真实类别为正,最后预测得到的结果却为负。TN(True Negative)表示样本的真实类别为负,最后预测得到的结果也为负。

ROC 曲线由负正类率(False Positive Rate,FPR)作为横坐标,正正类率(True Positive Rate,TPR)作为纵坐标。ROC 曲线距离参考线越远,其检验的准确度越高。AUC 是 ROC 曲线下的面积,其值越大越好。

对于不同的分析任务,可在上述指标中选择某几个作为衡量标准。例如,在疾病预测时,需要着重关注召回率,而不是精确率,因为疾病在多数情况下是正例(不患病),负例(患病)较少,两个类的样本比例差别很大的情况下,例如,100 条记录中,5 次发现患病,其中

4 次为误报，1 次为识别出来，相较于全部识别为正常的精确率 99%，虽然精确率降低为 96%，但是 Recall 却由原来的 0/1=0%上升到 1/1=100%，虽然误报了疾病（经过复查可以排除），但是却没有遗漏错过真正患病的人群。

可以通过分析软件对分类结果进行自动化分析，例如，在 SPSS Modeler 中可以在生成的模型后面连接一个“分析”节点，运行它即可获得前述的各项分析结果，其属性配置及分析结果如图 1.15 所示。

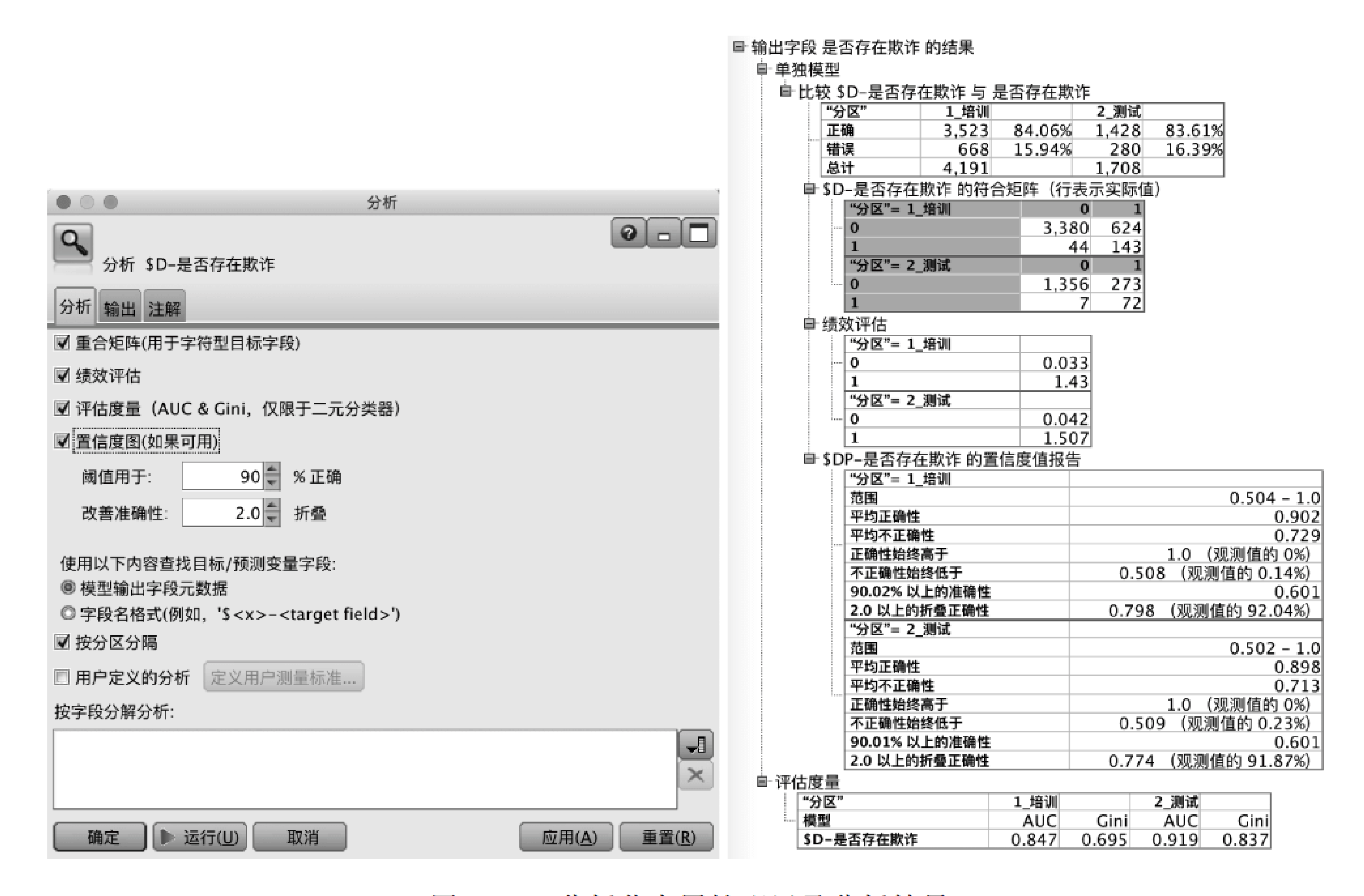


图 1.15　分析节点属性配置及分析结果

在属性选择中选中“重合矩阵”，可以显示混淆矩阵的数值，如果选中“置信度图”，则会显示置信度值报告，在评估度量中可以查看分区中训练集和测试集的 AUC 和 Gini 值。

1.6.2　聚类结果的评价

由于聚类是在没有类别标准的情况下对数据进行类簇划分，所以聚类分析结果的评价首先要由业务专家对其业务含义进行评估，通过应用到实际场景中来评价结果的好坏，看一下其区分程度。

应用散点图查看聚类结果，将聚类结果通过散点图的形式显示到二维或三维的空间中，查看各个聚类的分布情况，可以直观看到类与类之间的区分程度。例如，在 SPSS Modeler 中可以使用“图形板”节点可视化显示聚类中各维度变量的结果，除此之外，还有以下聚类指标。

1. RMSSTD（Root Mean Square STD）

RMSSTD 表示群体中所有变量的综合标准差，RMSSTD 越小表明群体内个体对象相似程度越高，聚类效果越好。

2. R Square

R Square 表示聚类后群体间差异的大小，R Square 越大表明不同的簇群间的相异度越高，聚类效果越好。

3. SRP(Semi Partial R square)

SRP 用于凝聚层次聚类算法的评价，表示当原来两个群体合并成新群体的时候，其所损失的群内相似性的比例。一般来说，SRP 越小，表明合成新的群体时，损失的群内相似性比例越小，新群体的相似性就越高，聚类效果就越好。

4. 簇类间距离

簇类间距离主要用于层次聚类算法的聚类评价，表示在要合并两个细分群体时，分别计算两个群体的中心，以求得两个群体的距离。一般情况下，聚类间的距离越小说明两个聚类越适合合并成一个新的聚类。

1.6.3 关联分析的评价

关联分析中几个重要的概念分别是支持度(Support)、置信度(Confidence)、提升度(Lift)。其中，支持度是指某一项集(若干个商品的集合)出现的可能性，即 support{x->y}$=P(x,y)$，如果支持度较低，则这一项集非频繁项集，不具有研究价值。

置信度是指项集中 x 出现的情况下，y 出现的概率，即包括 x 的项集中同时包括 y 的可能性：Confidence(x,y)$=P(y|x)=P(x,y)/P(x)$；提升度是在包含 y 的项集中，同时包含 x 的项集比例，Lift(x->y)$=P(y|x)/P(y)=$Confidence(x->y)$/P(y)$。提升度是为了弥补置信度的缺陷，主要用于分析 x 与 y 之间的关联强度，值越高说明关联性越强。

1.6.4 回归分析结果的评价

回归分析结果的评价分为两部分，首先是模型指标，是对模型结构合理性和显著性进行评价。其次是回归模型中回归系数的评价指标。

模型指标包括 R、R 方、调整 R 方(Adjusted R Square)、因变量预测标准误差(Std Error of the Estimate)、总离差、自由度、平均离差(Mean Square)、F 值、F 值的显著性水平(Sig)、模型个例数(N)，其中比较重要的是以下 5 个。

1. *R* 方

在模型概述表中查看，用于评价回归模型的总体表现，又称为确定性系数，表示自变量对因变量的解释程度，取值为 0～1，值越大，说明解释能力越强。

2. 调整 *R* 方

调整 R 方是对 R 方进行修正后的值，对非显著性变量给出惩罚，没有 R 方的统计学意义，与实际的样本的数值无关，相较于 R 方，其误差较少，是回归分析中重要的评价指标，其值越大说明模型效果越好。

3. 因变量预测标准误差

标识因变量的实际值与预测值的标准误差，其值越小说明模型的准确性越高，代表性越

强，拟合性越好。

4. *F* 值

在方差分析表中查看，用于检测回归方法的相关关系是否显著，如果显著性水平(Sig)指标大于0.05，表示相关性较弱，没有实际意义，如果Sig指标小于0.05，但是各自变量的Sig指标均超过0.05，就需要应用 t 检验来查看回归系数表中各变量的显著性水平，或者是由于自变量之间出现了共线性问题，需要通过逐步回归的方法将显著性较差的自变量剔除。

5. *N*

N 显示的是应用于模型的实际样本数量，可能有部分数据存在空值或其他异常值，导致模型的个案数少于样本数，如果发现其值较大，需要对数据重新进行预处理。

多元回归方程公式：

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e \tag{1.1}$$

要求每个 X_i 必须是相互独立的，其中 b_i 表示回归系数。回归系数可以从回归系数表中查看，其评价指标主要包括以下4个。

1) 非标准化系数(Unstandardized Coefficients)

非标准化系数就是多元回归方程式(1.1)中的 b_i，表现在几何上是斜率。由于其数值与实际的自变量数值的单位，彼此之间无法进行比较，为了对非标准化系数的准确性进行衡量，使用非标准化系数误差(SER)来对样本统计量的离散程度和误差进行衡量，也称为标准误差，它表示样本平均值作为总体平均估计值的准确度，SER值越小说明系数预测的准确性越高。

2) 标准化系数(Standardized Coefficients)

在多元回归分析中，由于各自变量的单位可能不一致，就难以看出哪一个自变量的权重较高，为了比较各自变量的相对重要性，将系数进行标准化处理，标准化系数大的自变量更重要。

3) t 检验及其显著性水平(Sig)

t 检验的值是以标准误差的单位度量观测样本统计量与假设值之间的差，t 值相对越大，表示模型能以更高的精度估计系数，其Sig/p指标小于0.05，说明显著性水平较高，如果 t 值较小且Sig/p指标较高，说明变量的系数难以确认，需要将其从自变量中剔除，然后继续进行分析。

4) B 的置信区间(95% Confidence Interval for B Upper/Lower Bound)

B 的置信区间用来检验 B 的显著性水平，主要为了弥补 t 检验和其Sig值的不足，如果 B 的置信区间下限和上限之间包含了0值，即下限小于0而上限大于0，则说明变量不显著。在SPSS分析时，可以选择“专家”选项卡中的输入选项进入高级统计，选中“参数估计”，以显示 B 的95%置信区间的上限和下限。

1.6.5　深度学习的评价

深度学习的模型也可以分为监督式学习和非监督式学习两种。用于分类的深度学习模型其结果的评价与分类算法一致，以准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1分值(F1 Score)为主，辅以ROC、AUC，并结合实际应用场景进行结果评价。

如果深度学习的应用方向是聚类的目的，数据源并没有进行标记，那么其模型结果的评

价按照聚类算法的标准来操作,如 RMSSTD、R Square、SRP 等。

1.7 数据分析团队的组建

随着大数据、人工智能广泛受到关注,各企业的决策者已经具备了数据驱动业务的意识,认识到数据分析对企业发展的潜在推动力,其中,在信息技术、金融等信息化程度较高的行业,数据分析团队建设处于领先地位在公共管理、医疗、能源、科教等领域中已经具备了信息化基础,也在逐步自建或外包数据分析团队,像制造业、建筑行业等传统行业还处在信息化建设时期,未来对数据分析的需求较大。

目前,数据分析团队属于新出现的职能部门,很多数据分析团队的建设过程也面临着一些问题,如数据分析结果很难落地、业务部门缺乏协作的动力、数据分析人才紧缺等,导致虽然公司领导对数据分析团队寄予厚望,但实际对业务带来的价值却有限。面对这些问题,就要求机构在组建数据分析团队时,要建立清晰的团队建设目标,将数据分析纳入决策流程,真正建立数据驱动的决策文化。

在实践中可按机构的信息化水平和业务特点渐进地构建数据分析团队。常见的数据团队的组织架构分为金字塔式和矩阵式,前者由首席数据官或项目经理作为领导者,带领数据科学家、数据工程师和业务专家,配合各个业务部门进行嵌入式分析工作,这种模式可以将分析技术进行复用,又可以快速响应业务部门的要求。矩阵式结构通常没有具体的负责人,而是以数据采集、数据清洗、数据分析、决策报告等工作来划分小团队,同一个小团队可以向多个业务部门提供服务,其好处是各数据小团队专业做自己擅长的技术,数据分析专业化程度较高,缺点是要求数据团队的成员对各业务部门知识都熟悉。

数据分析团队按照职能划分,可以分为项目经理、业务专家、数据提取人员、预处理人员、建模人员、测试人员,在实际的分析过程中可以将部分职能岗位进行细分或合并,如数据提取人员和数据预处理人员可为同一(组)人。

1.7.1 项目经理

项目经理或团队领导者通常肩负着定义团队目标、组建管理团队、出品数据分析报告等至关重要的职责,主要负责整个分析任务的目标设计、分工协调、方案设计和最终分析报告的总结生成等,其核心工作在于将各职能人员的目标尽可能保持一致,并对各成员的输出进行确认,以防出现数据处理不合格影响模型的效果,最终无法得到最优模型。

要求项目经理具备丰富的项目管理经验,对算法、模型、技术有一定的了解,最好是技术出身,既可研究技术,又可沟通业务,能够与业务部门合作,减少团队成员的工作阻力,激发团队热情,挖掘更多的数据价值。

1.7.2 业务专家

在某些专业化较强的领域中,数据分析人员需要尽快熟悉业务需求,在业务专家的指导下对需求或目标进行细化,以制定相应的数据要求说明书和分析模型设计规划。业务专家的角色在数据分析中非常重要,对模型在实际应用中进行应用检验都需要他们的确认,否则

模型容易出现某些行业常识性错误。

1.7.3 数据工程师

数据工程师须具有编程能力,对算法、数据架构、软件工程有深入理解,如果对数据分析有一定的理解更好,其主要工作是将分析模型集成和应用,除此之外,还要对数据进行收集、整理和数据清洗,好的数据质量可以极大减少建模的工作量和提高模型的性能。另外,模型在实际业务流程中的部署和维护都需要工程师有较高的软件系统设计能力和开发能力。从职能上可将数据分析工程师细分为:数据平台架构师、开发工程师、运维工程师等。

在数据分析过程中,很多数据是可遇不可求的,在实际分析过程中需要对第三方的数据进行提取,以补充到数据集中,要求这部分人员有一定的编程经验,特别是要掌握一定的爬虫技术,对 HTTP 等网络协议有一定了解,能够在较短的时间内编写相应的代码对网站内容进行爬取。常见的爬取编程语言为 R 语言、Python 等,其优势是目前有较多的第三方框架支持快速抓取内容,当然,Java 或 C# 也可以实现相同的功能。

数据预处理的主要工作是对数据进行数据清洗,包括去除空值、异常数据,从而提高原始数据集的质量,另外一项工作是通过对数据进行多表关联查询和统计,将复杂字段统计之后提交给模型分析人员,减少模型的预处理时间,提高效率,并可以在建模之前对数据进行探索,能够进行统计型的数据分析。

1.7.4 数据建模人员

数据建模人员包括两大类,分别是数据分析师和数据挖掘工程师,前者要有科研能力,主要工作是对行业数据进行整理、分析,以做出行业研究、评估和预测等,通过使用工具软件来实现数据的商业意义。数据分析师至少要熟练掌握 SPSS、Statistic、Eview、SAS 等数据分析软件中的一种,最好具有一定的编程能力。

数据挖掘工程师需要具有一定的数学知识,掌握类似高等数学、概率统计、线性代数等数理常识,要对各种分类、聚类、关联、回归等算法特点和应用条件较熟悉,能够结合业务情况和实际提供的数据集进行算法选择,并且能够对算法进行一定程度的调优。

1.7.5 可视化人员

一图胜千言,分析结果的呈现是整个分析任务的整体表现。好的数据可视化不仅仅采用图形表格,而且将数据变化的过程和趋势进行动态展示,需要可视化人员依据行业或产品进行设计,按照场景和性能要求,选择合适的可视化技术,并制作样例。优秀的可视化工程师不仅可进行视觉设计,还具有一定的前端开发能力,使用 Node.js 或其他第三方组件进行数据动态展现。

1.7.6 评估人员

模型建好后,需要在测试环境和生产环境中进行测试和验证,评估人员在业务专家的配合下对模型进行不同应用场景的测试,以便查找模型中的过拟合、异常情况处理不足等问题,特别是在医疗领域,需要经过多轮反复验证后才可以投入使用。

1.8 数据分析人才培养的难题

数据分析行业可附加至其他行业中，为各行各业提供技术支持，所以这方面的人才需求缺口很大，只要具有一定的数据分析能力，薪资待遇普遍较高，但是岗位要求不低，需要的是复合型人才，具有发现问题、分析问题、解决问题的能力，能够结合商业、数据、问题等形成解决方案。

具体来看，数据分析人员需要掌握数据挖掘、统计学、数学等基本的数理原理和常识，需要掌握并熟练运用某一数据挖掘软件，如 SPSS、SAS、R 等，除此之外，还需要熟悉各类模型算法的特点，以及在各种场景中如何进行选择和应用，相应的人才标准较高，培养难度较大，需要经过实战案例训练逐步提高数据挖掘水平。

1.8.1 数理要求高

鉴于数学相关专业的学习曲线较为陡峭，大多数人对于数学相关的理论望而生畏，越发难以深入学习，目前对从事数据分析行业的人才这方面的要求较高，在数据分析过程中需要应用高等数学、线性代数、概率论、离散数学、统计学等，对数理理论缺乏原理上的研究，在模型建模过程中很难做到创新，只能照猫画虎进行模仿。

1.8.2 跨学科综合能力

如果是开发人员，可以通过编程实现，可以使用 Python 等语言应用相关模型，或者使用 Weka 框架来实现，这就需要有一定的软件工程师的背景，或者具有较快的跨学科学习和应用能力，可以快速使用现有框架进行模型建模和应用。

在目前的软件从业人员中，大部分开发人员对数理知识并不精通，特别是统计学等理论，而数学、统计学等专业人员往往更精通理论，而缺少编程经验，对于快速实现模型的应用又具有局限，特别是在数据提取、预处理、分析结论可视化等方面，需要与软件开发进行配合。

数据分析过程中需要掌握的技术除了 SPSS 等建模软件和分类、聚类、回归等算法外，还需要对 Hadoop、Spark、Storm、MapReduce 等平台具有应用经验，对编程语言的要求是至少熟练运用 C++、Java、Python、R 等语言中的一种，同时还要求熟悉数据库、存储等知识，具有一定的数据优化能力。综合能力要求较高，而上述技术或框架近几年刚开始流行且更新很快，每个分支达到熟练应用均需花费较长时间进行学习与实践，对从业者能力和能否持续学习都需要考验。

1.8.3 国内技术资料少

由于数据分析属于 IT 行业新兴行业分支，国内的技术资料较少，如果要与时俱进，须直接阅读国外资料，这就要求具有一定的英文水平，能够流畅阅读国外技术资料和书籍，同时要具有较强的信息检索和查找能力，遇到问题时，可快速定位问题的原因，并获取其他人的解决方案。

1.8.4 实践机会少

目前数据分析行业的实践机会较少，一方面是企业对数据分析的投入相较信息化建设较少，数据分析项目虽然越来越多，但总体数量上仍然具有更大的潜力；另一方面，软件开发和数学专业的从业人员更愿意停留于当前专业领域中，对于主动从事跨专业研究的动力不足，随着数据分析人员的需求增多，待遇随之水涨船高，必将吸引更多的人才进入数据分析行业。

数据分析行业虽然前景好、待遇高、人才需求大，与其他行业一样，并非所有人都适合从事此行业，入行前首先要对岗位和自身进行评估，好好思考这些问题：What，Why，How，即：数据分析行业是干什么的？有哪些知识要求？我为什么要加入这一行业？是因为兴趣吗？我自身有哪些优势条件？要想达到较高的水平，要如何干？可从以下几个方面进行评估。

职业爱好：数据分析行业仍然属于IT行业，这一行业普遍要求务实、严谨、少说多做的风格，属于在后台默默工作付出的那一层级，需要思考能否与枯燥的代码为伴，并乐在其中。

思维能力：数据分析人员要求具有较强的逻辑思维和推理能力，需要从数字中探寻出业务的核心规律，最好能有见微知著和创新的能力，如果经过培训之后仍然对数据无感觉或不敏感，可能说明不适合与数据打交道。

学习能力：技术发展很快，需要不断学习新的技术、新的处理过程等，这是与其他行业差别较多的地方。在IT行业中，某一项技术从流行到消失一般只有几年的时间，所以要求从业人员不断学习，不断提高。当然，IT行业的原理性知识，如数理知识、数据结构、操作系统等技术理论变化很少，主要的变化还是理论的具体应用，但万变不离其宗。

沟通能力：数据分析行业需要跨部门沟通，与业务部门、研发部门进行合作，特别是项目经理等领导岗位，既要有合作意识，又要有推动能力，在协调过程中争取更多的支持，减少摩擦，使最终分析结果能够给各企业带来正向收益。

业务知识：理解业务知识可以快速选择合适的模型和算法，少走很多弯路，不需要对模型结果反复评估，就可以确认此模型是否符合业务需要。理解数据与业务流程、组织架构对企业的影响，对业务具有敏感度可以更好地推动数据分析为产品服务，不至于闭门造车，最终帮业务部门提供快速决策支持。

第 2 章

数据挖掘算法的选择——保险产品推荐

数据挖掘算法没有好坏，每种算法都有一定的适用范围。数据分析师可以根据数据以及数据分析需求的特点，大致选择几种方法，然后通过实验比较确定合适的挖掘算法，并逐步调优。

按照数据挖掘建模标准（CRISP-DM）的流程，首先要定义商业问题，理解业务背景，对业务需求有基本的了解，然后对相关的数据进行探索、预处理，分析其特点，进而确定几个可能的模型，并对其进行验证评估，最后选择分析结果较优的算法对其进一步调优，使结果尽量解决客户的问题，最后将模型进行应用部署。

上述流程中因实际数据挖掘中的任务目标和数据特征千差万别，像数据预处理等可忽略，但流程中任何一步出现问题，构建出来的模型可能就会毫无应用价值。为了说明数据挖掘算法的选择过程，现在以保险数据分析为背景，讨论数据挖掘算法选择的一般方法。

2.1 业务理解

从商业的角度对业务部门的需求进行理解，包括商业背景分析、理解行业术语、业务成功标准、企业需求和设想等，对业务不了解，在模型选择上容易走弯路，而且容易陷入细节，虽然模型准确且合理，但业务用户觉得一文不值。

保险行业具有“避税”“避债”“可继承”的特点，成为高净值人士的青睐之选，特别是目前国内的投资机会较少，股市、楼市、实体经济等风险较高的情况下，保险业迎来蓬勃发展的春天，大量保险企业纷纷推出各种各样的人寿型、医疗型、投资理财型等，竞争激烈。

众所周知，由于保险业务各项条款细则较复杂，且出险认定以及赔付流程较长，传统保险采用代理人制度，即被保险人通过保险的代理销售人员购买保险，由代理人负责条款解释说明、订单确认和购买，以及出险的赔付等，受限于代理人的能力和业绩考核压力等，保险代

理人并不能有效地向客户推荐其需要的险种，反而使客户对保险代理人和保险产生抵触心理，影响了企业的品牌信誉。因此，目前保险行业需要数据挖掘技术支持，通过对客户过往的保险购买记录分析客户特点，并以此为依据验证是否需要向其推荐其他险种，对于分析结果中某一客户购买概率较低的险种，则不再向其推荐，不仅减少了资源浪费，而且提高了投放精准性，促进保险公司的业务发展。

在商业中，最关键的是提炼问题，确定要解决什么问题，确定业务目标是战略问题，而选择和确认模型是战术问题。很多计算机专业的人才具有很好的解决问题能力，但缺少在大量数据集中寻找出问题、总结商业模式等能力，提出业务上要解决的问题和确认分析任务目标在数据分析过程中至关重要。目前，大部分业务目标主要是业务部门提出需求，但业务部门在面对海量数据时，只能提出一些在其技术认知范围内的直接问题，会有较大的局限，高层次的数据分析师不仅要熟知数据分析技术，还要了解商业及其他领域基础知识，能够帮助客户从数据中挖掘出新的商业模型或商业机会。

本例中，保险公司提供了以家庭为单位的历史保险投保记录，同时给出了家庭及其成员的各种属性统计结果，总共 86 个字段。保险公司目前正准备向客户推荐一款房车险，希望通过对这些保单记录和属性信息进行挖掘，分析哪一类客户倾向于购买此保险，并希望了解分析的过程和原因。以上目标比较简单、直接，就是要找出移动房车险客户的特征，然后依据这些特征在客户库中有选择性地进行营销活动，提高销售效率，减少运营成本。

获取某一类客户的特征后，就可以在后续的保险推广中应用相应规则，减少大量低效打扰客户的病毒式推广，这个业务目标在企业经营活动中具有很强的普遍意义，在其他行业中也有很多类似的情况。例如，酿酒企业想了解哪类客户更愿意购买新出品的一款红酒，或者车企想要知道推出某款新车的受欢迎程度等。

2.2 数据分析目标

提出业务目标后，要将其转化为数据分析的目标，即将商业问题转化为数学问题或数据分析问题，首先要具有一定的行业领域知识，了解行业的痛点，同时对能拿到的数据进行分析，通过思维发散，提出各种各样的算法模型，最后将问题简化，变成纯粹的数据挖掘问题。例如，商业目标是希望提高转化率和缩短转换周期，首先整理数据，获得客户信息等静态数据和客户操作记录的动态数据，然后研究转化成功的用户具有哪些特征，总结规律，提出模型，从而改善营销方式和优化购买流程，提高投资回报率。

本例中的数据集中存在大量的客户属性数据和保险购买的统计数据，需要提出某一规则或算法，将客户的属性信息、购买记录输入模型，由模型给出一个是否购买的结果值。可以看作是分类问题，将客户分为购买房车险和不购买房车险两个类别；也可以进行关联分析，将客户购买保险作为规则项，还可以对样本进行聚类分析，由于业务目标中需要知道某些规则，对于数据分析的方法而言，样本数据特征各异、数据量和样本中分布情况也有较大差别，所以无固定模式可以直接应用，在分析过程中先对数据进行探索，然后粗选某些分析算法，再逐渐调优解决业务问题。

2.3 数据探索

对数据进行理解以找出问题的影响因素，主要包括数据质量检查、描述性数据统计、探查各变量的意义及其相关关系、验证其中隐藏的信息和知识，对数据不理解，选择数据时容易出现覆盖不全、不完整、错误数据等问题，建模结果就会片面或不稳定，甚至出错。

对数据进行探查，以发现其主要特点，理解数据结构和各变量的意义，对数据形成直观认识，包括单变量的分布情况分析、多变量关系分析等，在探索过程中可以应用可视化技术从中看出某些规律，如散点图、箱图、直方图等。

本例中的数据为保险公司统计后的结果数据，是以家庭为单位将客户属性、购买保险的种类、金额等数据进行计算，以家庭房产数为例，统计整个家庭中房产的总数量，结果的取值范围为1～10，即最少1套，最多10套，而家庭中宗教情况是按照家庭成员信仰某一宗教的人数占总人数的比例来统计，在提供的数据中，其值范围为0～9，分别表示0、1%～10%、11%～23%、24%～36%、37%～49%、50%～62%、63%～75%、76%～88%、89%～99%、100%。不同的维度字段，其中数值代表的意义不同，需要根据维度的含义和意义进行具体分析，按照字段数值的类型划分，可将数值分成5个类别，分别是实际数值型、L0、L1、L2、L3、L4，其详细说明见表2.1。

表2.1 变量取值类型说明

字段类型	字段类型说明
实际数值型	家庭房产数量(1～10)、平均房产数量(1～6)
L0	客户子类别标签，取值为1～41，代表高收入、单身青年、中产阶级、丁克等
L1	年龄范围，取值为1～6，1表示20～30岁，…，6表示70～80岁
L2	客户主类别标签，取值为1～10，分别代表功成享受、退休信教、保守家庭等
L3	百分比，0～9，0表示0，1表示1%～10%，…，9表示100%
L4	金额(欧元)，0～9，0表示0，1表示1～49，2表示50～99，9表示超过20 000

总的记录数为5822条客户数据，每条记录包括86个变量，前43个变量为人口属性，是基于邮局系统中的门牌号来统计的每户家庭中各成员的信息，然后进行合并计算将结果作为最后属性的结果；后面的变量为产品购买属性，即之前购买过哪些保险；最后第86个字段是目标字段，表示是否购买移动房车险，取值为0或1，即分析客户是否会购买这一险种。

2.3.1 数据质量评估

样本数据的质量直接决定了最终模型的准确性，高质量的数据覆盖了模型需要的各种情况，且能够如实反映除模型训练之外的所有数据，但是这样的样本数据往往可遇不可求，在实际分析中受业务系统等限制，难以将数据收集完整，总会存在各种各样的问题，如出现样本不平衡，重要数据无法提供，存在错误数据。“垃圾进，垃圾出”，以此建立的模型必然无法应用，对于样本数据质量的评估显得尤其重要。

在本例中，由于给定的数据包括86维数据，如果对每一变量进行单独分析，耗时较长，为了快速查看各维度数据的基本特点，可以应用IBM SPSS Modeler中的数据审核结果对

数据进行评估。Weka 等分析软件中也具有此类功能,除此之外,还可以应用直方图、散点图等对某一字段进行自定义分布间隔来查看样本分布。

从数据的准确性、完整性、一致性等维度找出样本存在问题,图 2.1 是数据审核节点的结果,可以看到各自变量的类型和数据分布情况,以及极值、平均值、标准差、偏度、类别数(非连续型变量)、有效的记录数。其中大多数变量呈现正态分布,部分自变量为偏正态分布,还有部分变量分布没有规律,比较散乱。

文件(F)　编辑(E)　生成(G)

审核　质量　注解

字段	样本图形	测量	最小值	最大值	平均值	标准差	偏度	唯一	有效
客户次类别		名义	1.000	41.000	--	--	--	40	5822
每房人数		连续	1.000	5.000	2.679	0.790	0.183	--	5822
客户主类别		名义	1.000	10.000	--	--	--	10	5822
新教比例		标记	0.000	9.000	--	--	--	10	5822
其它宗教比例		标记	0.000	5.000	--	--	--	6	5822
无宗教比例		标记	0.000	9.000	--	--	--	10	5822
已婚占比		标记	0.000	9.000	--	--	--	10	5822
其它关系占比		标记	0.000	9.000	--	--	--	10	5822
单身占比		标记	0.000	9.000	--	--	--	10	5822
高等教育		标记	0.000	9.000	--	--	--	10	5822
低等教育		标记	0.000	9.000	--	--	--	10	5822
高管		标记	0.000	9.000	--	--	--	10	5822

图 2.1　数据审核节点的结果

通常,正态分布的样本更符合预期,说明其样本数据分布较合理,覆盖了大多数的情况。当然,样本为偏正态分布时可以应用对数、倒数、指数等变换将其转换为正态分布,从而改进模型分类结果的准确性。

在"质量"选项卡中查看数据质量,如图 2.2 所示,可以看到完整字段和完整记录的比例,以及空值、字符型空值、离群值、极值的数量等,可以看到样本数据中没有上述异常数据,样本所有字段的数据完整度均为 100%。

由于上述工具与业务无关,它们并不能检查业务数据存在的问题,所以在分析过程中需要利用业务知识查看数据的合法性和准确性,以及是否存在异常值,这些工作需要从业人员认真查看数据的实际取值,而非仅仅看通用的质量指标,可应用箱图等图形对有疑问的字段数据进行详细探查,具有一定编程或数据库使用经验的人员可以直接操作样本库查找异常记录,并按照实际需要决定是否将其剔除出训练集或对其进行修正。

审核 质量 注解

完整字段(%): 100%　完整记录(%): 100%

字段	测量	离群值	极值	操作	缺失插补	方法	完成百分比	有效记录	空值	字符型空值	空白	空白值
客户次类别	名义	--	--	--	从不	固定	100	5822	0	0	0	0
每房人数	连续	0	0	无	从不	固定	100	5822	0	0	0	0
客户主类别	名义	--	--	--	从不	固定	100	5822	0	0	0	0
新教比例	标记	--	--	--	从不	固定	100	5822	0	0	0	0
其它宗教比例	标记	--	--	--	从不	固定	100	5822	0	0	0	0
无宗教比例	标记	--	--	--	从不	固定	100	5822	0	0	0	0
已婚占比	标记	--	--	--	从不	固定	100	5822	0	0	0	0
其它关系占比	标记	--	--	--	从不	固定	100	5822	0	0	0	0
单身占比	标记	--	--	--	从不	固定	100	5822	0	0	0	0
高等教育	标记	--	--	--	从不	固定	100	5822	0	0	0	0
低等教育	标记	--	--	--	从不	固定	100	5822	0	0	0	0
高管	标记	--	--	--	从不	固定	100	5822	0	0	0	0
农场主	标记	--	--	--	从不	固定	100	5822	0	0	0	0
中层管理者	标记	--	--	--	从不	固定	100	5822	0	0	0	0
技术工人	标记	--	--	--	从不	固定	100	5822	0	0	0	0
非熟练劳工	标记	--	--	--	从不	固定	100	5822	0	0	0	0
社会阶层A	标记	--	--	--	从不	固定	100	5822	0	0	0	0
社会阶层C	标记	--	--	--	从不	固定	100	5822	0	0	0	0
社会阶层D	标记	--	--	--	从不	固定	100	5822	0	0	0	0
租房子	标记	--	--	--	从不	固定	100	5822	0	0	0	0
房主	标记	--	--	--	从不	固定	100	5822	0	0	0	0
一辆车	标记	--	--	--	从不	固定	100	5822	0	0	0	0
无车	标记	--	--	--	从不	固定	100	5822	0	0	0	0
公共社保	标记	--	--	--	从不	固定	100	5822	0	0	0	0
私人社保	标记	--	--	--	从不	固定	100	5822	0	0	0	0
收入低于30	标记	--	--	--	从不	固定	100	5822	0	0	0	0
收入45-75	标记	--	--	--	从不	固定	100	5822	0	0	0	0
收入75-122	标记	--	--	--	从不	固定	100	5822	0	0	0	0
平均收入	标记	--	--	--	从不	固定	100	5822	0	0	0	0
购买力水平	标记	--	--	--	从不	固定	100	5822	0	0	0	0
个人第三方保险	标记	--	--	--	从不	固定	100	5822	0	0	0	0
投保车险	标记	--	--	--	从不	固定	100	5822	0	0	0	0
投保火险	标记	--	--	--	从不	固定	100	5822	0	0	0	0
第三方私人险数量	标记	--	--	--	从不	固定	100	5822	0	0	0	0
投保车险数量	标记	--	--	--	从不	固定	100	5822	0	0	0	0
投保火险数量	标记	--	--	--	从不	固定	100	5822	0	0	0	0
移动房车险数量	标记	--	--	--	从不	固定	100	5822	0	0	0	0
分区	名义	--	--	--	从不	固定	100	5822	0	0	0	0

图 2.2　数据审核节点中数据质量结果

2.3.2　探索数据统计特性

描述性统计分析是用统计学的指标来描述数据特征的一种方法，其理论基础是数理统计学知识，主要包括数据的集中趋势、离散趋势、数据分布等特征，它是数据分析的第一步，也是进一步分析的基础。描述集中趋势的指标有均值、众数、中位数等，描述离散趋势的指标有极差、方差、标准差、四分位等，描述数据分布情况的指标有偏度、峰度等，前者是对数据分布对称性的描述，后者是对数据分布平峰或尖峰程度的描述，主要用于查看数据是否符合正态分布。

经过描述性统计分析之后，就可以有针对性地分析其中部分字段，分析样本中包含的某些特点，一方面可以验证数据的质量，此外也可以对样本数据有更加直观的感觉，同时作为模型结果的验证也非常有用。

目前很多分析软件或模块都有统计分析功能，例如在 SPSS Modeler 中可以从“节点选项板”中的“图形”选项卡中选择合适的图形节点对数据进行统计分析，而在 Python 中可以先使用 NumPy 和 SciPy 进行统计分析，然后用 Matplotlib 工具库来可视化显示结果。

本例中是为了查找投保移动房车险的家庭特征，由于移动房车险的类别为投(值为 1)或不投(值为 0)两种情况，所以可应用 SPSS Modeler 中的箱图来查看不同的变量对因变量的区分度，结果如图 2.3 所示。

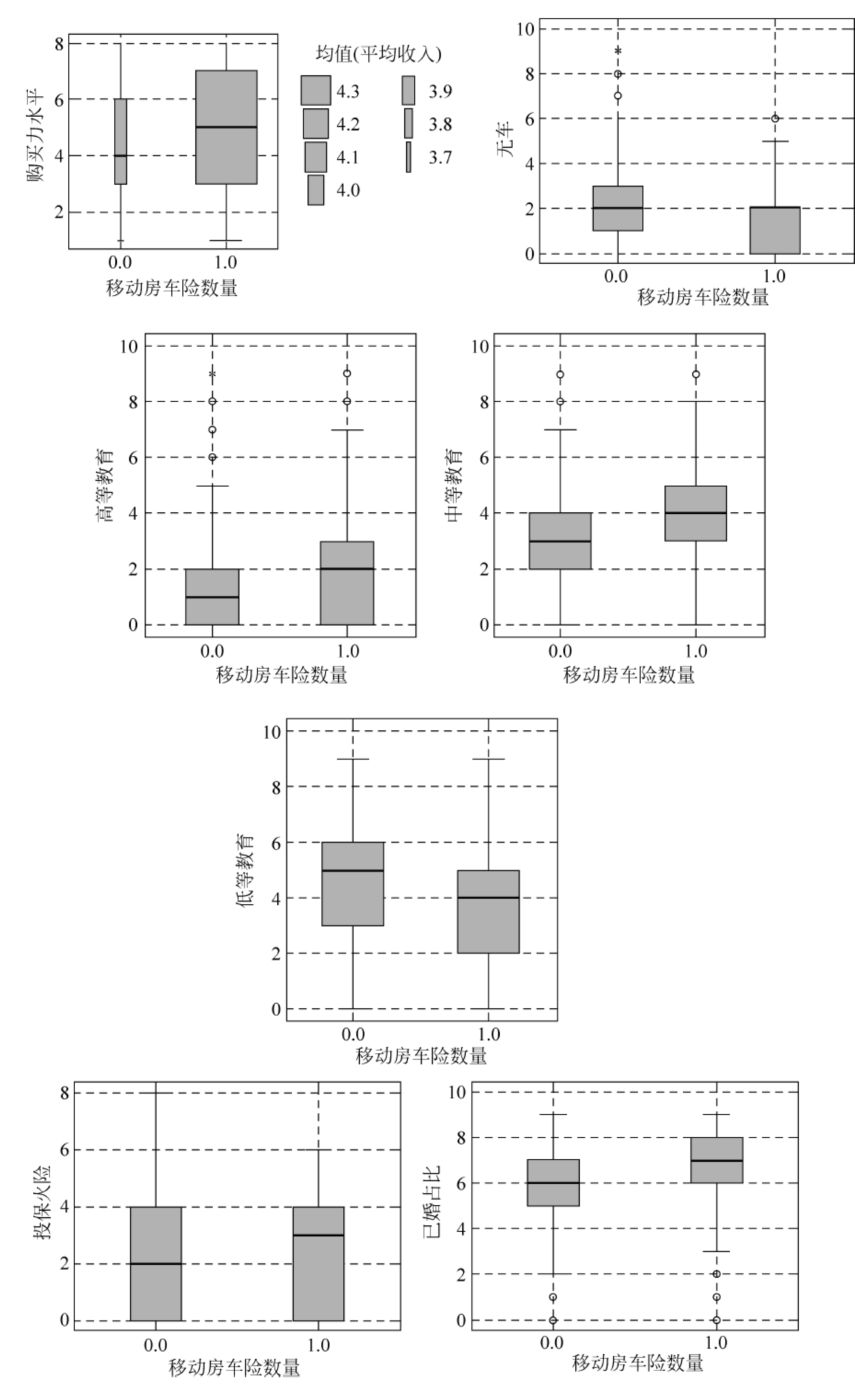


图 2.3　购买移动房车险家庭统计分析

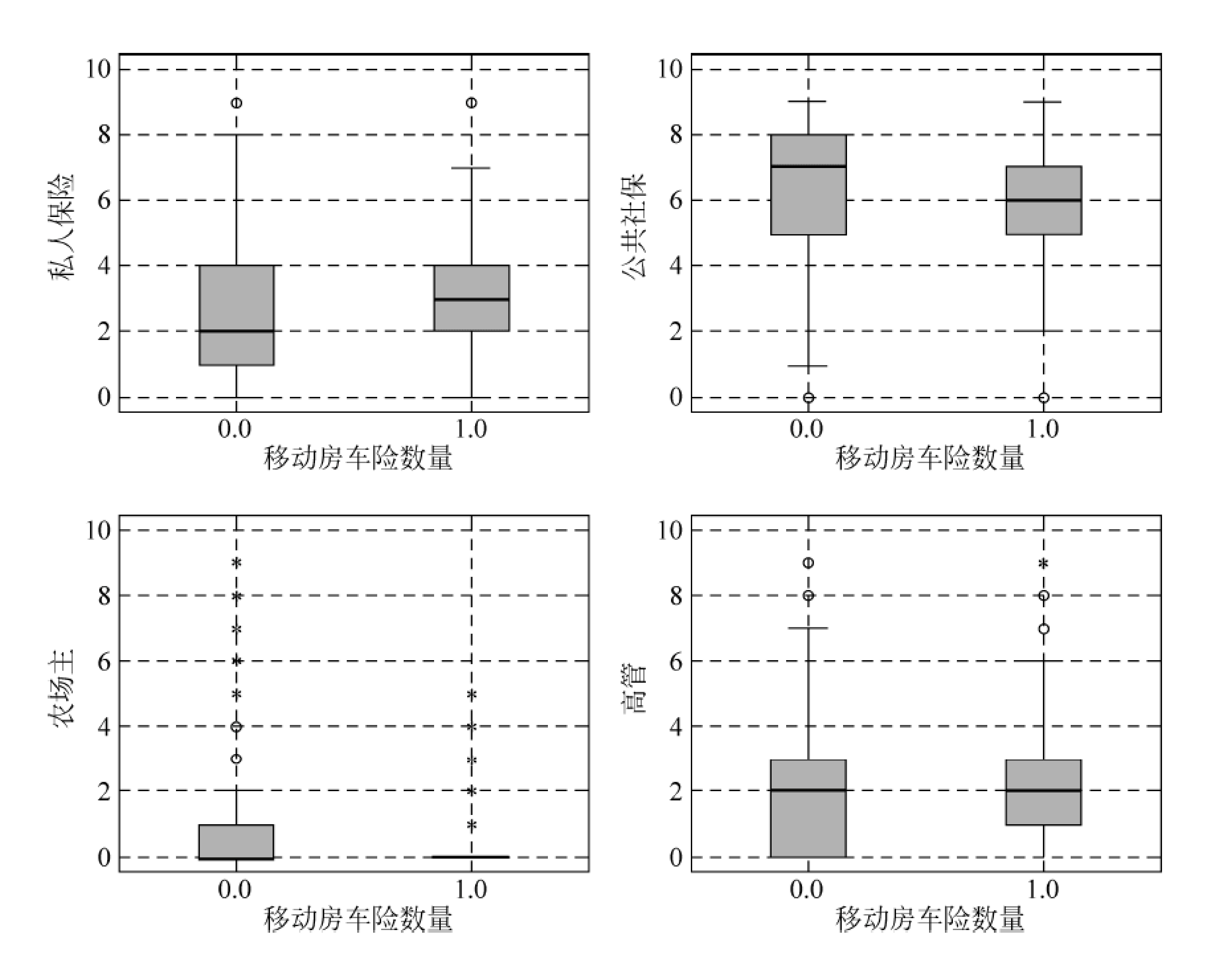


图 2.3 （续）

经过比较，发现购买房车险的家庭相较未购买此保险的家庭具有以下特征：

（1）购买力水平较高，平均收入较高。

（2）平均教育水平较高。

（3）投保火险的比例略高。

（4）家庭成员中已婚的比例较高。

（5）私人保险投保比例较高。

（6）公共社保的投保比例较低。

（7）农场主这类人群极少投移动房车险。

（8）高管层次的人群比例较高。

从这些特点中可以得出初步的结论，投保移动房车险的家庭其经济实力明显较强，教育程度较高，社会地位较高，保险意识和理念较强，基本上为中产阶级及以上人群，所以这个险种的目标人群可以初步进行定位。由于此结论相对模糊，所以还需要应用分析模型进行详细分析，对结果进行量化，形成可操作和可应用部署的模型算法。

2.3.3 数据降维

降维是一种常见的数据预处理手段，一般情况下，样本的字段数较多，特别是在大型系统中，由于各不同应用方向的信息系统记录的数据种类很多，经过综合之后就容易产生维度灾难，使得在数据分析过程中模型训练时间超长，且冗余字段也影响模型的准确性，易产生

误差，所以，在大多数情况下要对字段进行降维处理，将对模型结果影响不大的字段剔除，或者将其进行变换后再输入至模型中。

在数据分析中可以使用分析软件附带的字段重要性评估模块来实现，也可以通过逻辑回归等模型进行评估，由模型给出显著性变量，如果变量对模型的贡献较少，可以考虑将其剔除。

本例中数据维数较多，需要进行降维处理，在 SPSS Modeler 中选择“特征选择”节点，如图 2.4 所示，目标设置为“移动房车险数量”，即家庭中投此保险的数量，输入变量选择其他所有字段，并使用分区字段，其他选项采用默认设置。

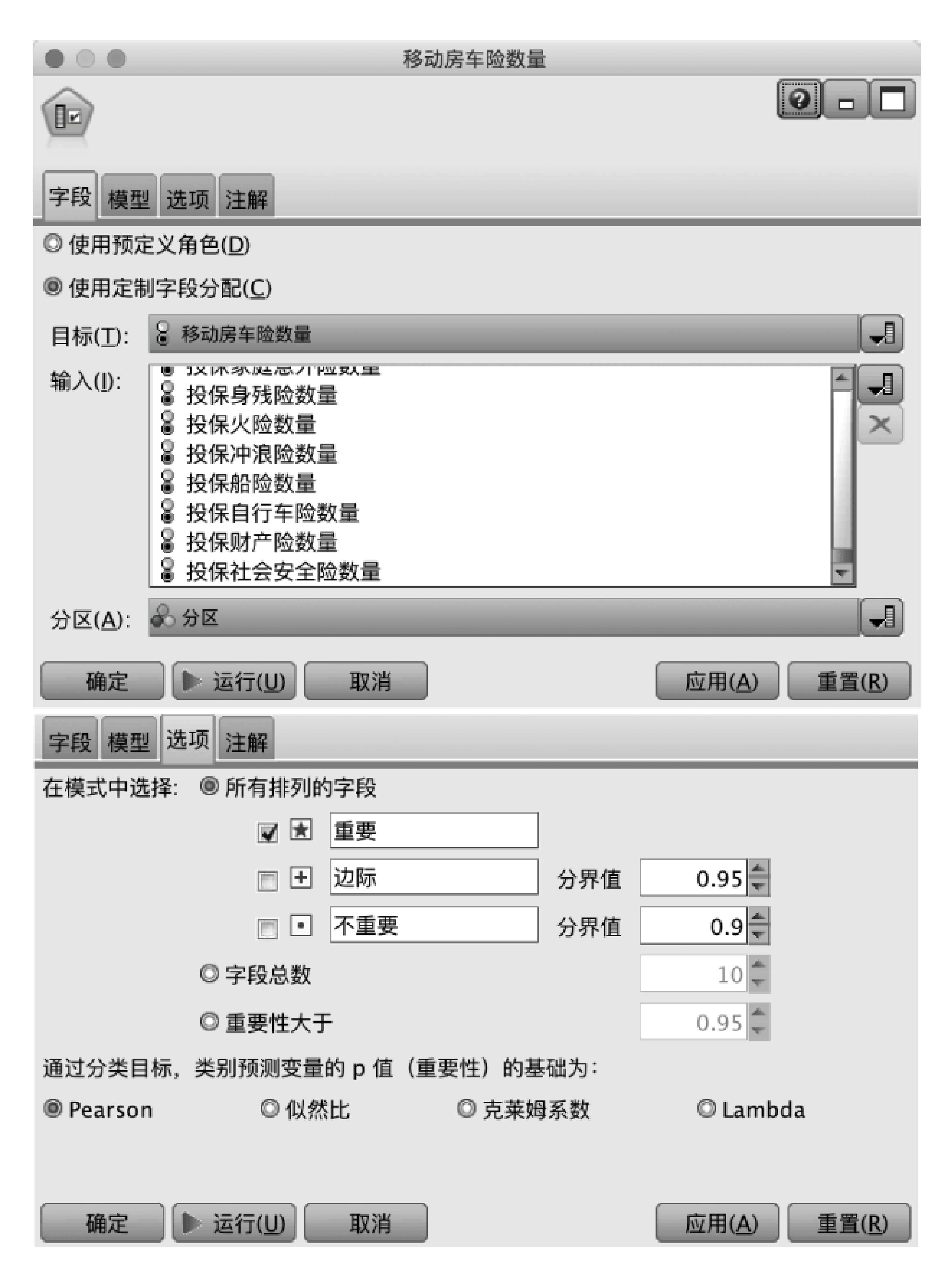


图 2.4　数据特征选择

运行模型后的结果如图 2.5 所示，结果中按重要性依次列出 36 个重要变量，此外一个边际变量，其他为不重要变量，这些变量并非不起任何作用，而是对结果的影响较小，在后续模型选择过程中也可以将其作为输入变量进行分析。

不同的业务需求和分析目标中，字段重要性等级的分界值不同，由于当前场景中字段较多，边际分界值选为 0.95 比较合适，而自变量较少时可以适当降低分界值，或依据实际业务场景进行设置。



图 2.5 特征选择结果

2.4 模型选择过程

在模型选择过程中，首先按照任务目标的要求和数据特点提出多个可能的模型，然后对这些模型进行详细分析，并选择具有较好区分效果的算法模型进行参数优化。自变量和因变量中的字段值大部分为分类类型，这种特点决定了比较适合应用分类算法，而分类算法种类较多，如果每种算法都进行验证，工作量较大，一般先采用自动分类技术筛选几类算法，然后再逐步确认。

模型选择要符合业务应用场景，很多人认为模型只要做到可预测就是一个好的模型，不符合业务需求，再好的模型也没有意义。在实践中，如果在业务上要求模型具有较强的可解释性，就不适合应用类似神经网络等黑箱模型，最好采用像逻辑回归、岭回归、决策树等可量化解释的模型，结果可量化且可验证。

模型的预测能力是指其泛化能力，在新的独立测试样本上的预测能力，多数情况下，随着模型复杂度的增加，其在训练集上的泛化能力增强，但在测试集上测试误差变大，甚至会出现过拟合现象，而模型选择阶段的主要目标是获得泛化能力最好，同时也是最稳定的模型。预测能力的评估除了从统计的角度验证外，还要从业务的角度进行验证，确认其在商业问题上的支撑力度。

在模型选择和评价过程中，一般将数据分为训练集、验证集、测试集 3 个不相交的集合，也可以将训练集和验证集合为训练集。总之，训练集和验证集是为了做模型选择，测试集是

做模型评价，其中测试集是独立的。当然，如果反复使用同一个测试集进行模型验证，也容易将其转化为验证集，产生过拟合，所以在实践中可将测试集分为多个独立的集合。

模型选择的核心思想是从众多可选模型中选择最佳的模型。例如，使用重复抽样对训练集进行划分，构建多个训练集和验证集，然后分别计算其近似测试误差，代表方法为交叉验证（Cross Validation，CV）和自助法（Bootstrap），这类方法需要较多的计算时间。另外，也可以基于似然来建立单一训练集误差和测试误差之间的关系，由于更加直接，所以计算速度更快，但结果可能不及理论预期，代表方法有AIC、BIC等。随着目前计算机处理能力的大幅提升，还是推荐使用简单可靠的CV法或GCV法来做模型选择。

图2.6是移动房车险投保情况分析流程，其中包括3部分：数据预处理、模型筛选与选择、模型优化。数据预处理在前面已经分析过，这里主要讨论模型的选择，在数据挖掘过程中，很难一次性确定哪个模型效果较好，所以模型的选择变得尤为关键。

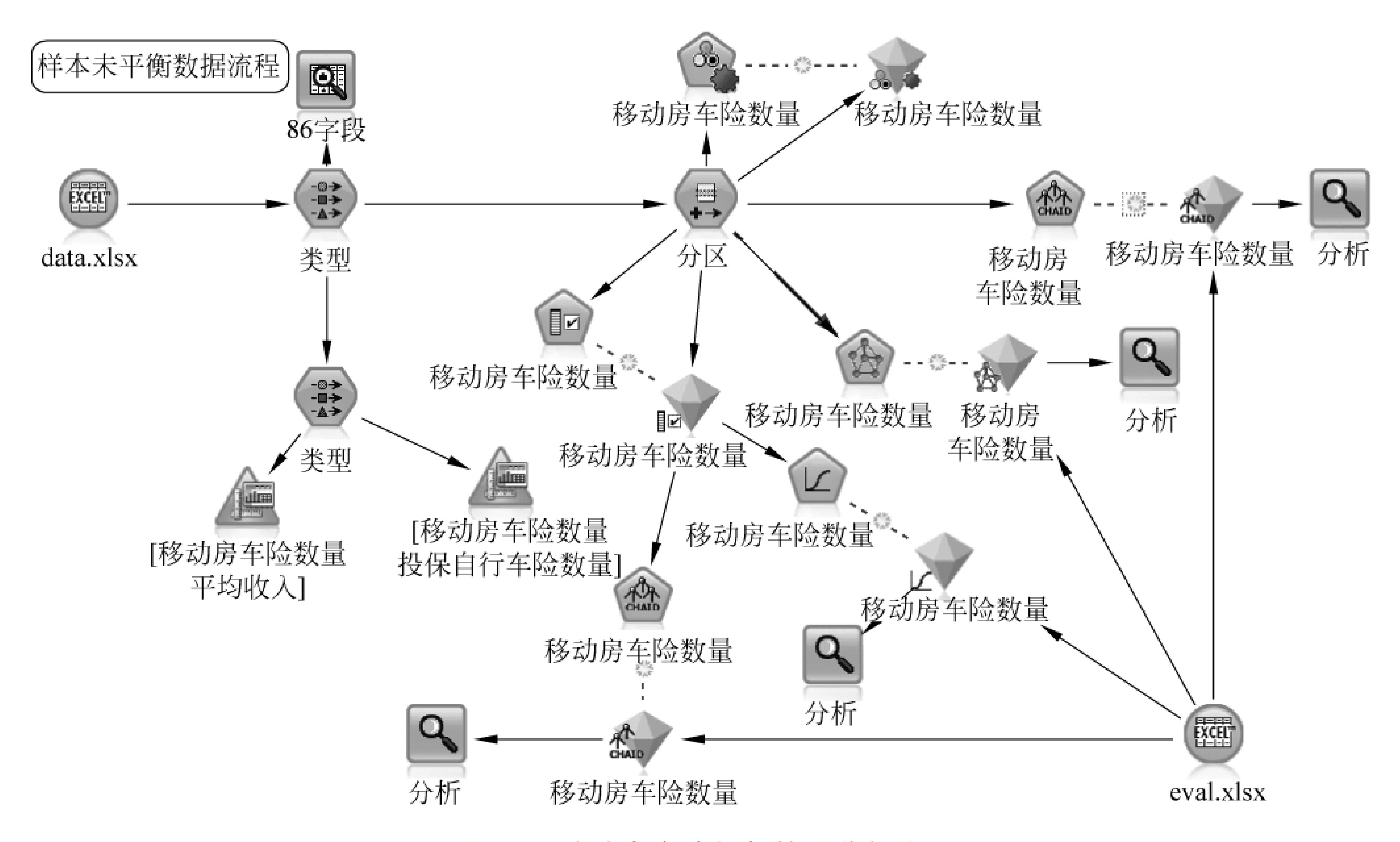


图2.6　移动房车险投保情况分析流程

如果软件具有自动算法选择功能，可以用它来初选几种算法，然后查看各模型的结果，并对其进行评估。一般情况下，自动算法并不能完全代替人工选择，需人工干预使模型可选范围逐渐缩小，最后选择一种或少数几种算法进行优化和比较。此外，如果数据分析工具未提供自动算法选择功能，可以根据数据特点以及分析问题初步选择几个模型进行测试，一方面可对数据特征进一步加深理解，另一方面也可作为“试金石”逐步扩展应用其他类别的算法。总之，算法的选择过程是一个不断地尝试、失败、再尝试的过程，最终发现数据中的规律。

2.4.1　算法初选

很多数据分析软件中都具有自动算法选择的模块，用于帮助用户以默认模型参数值来运行所有模型，然后按照某一评估准则筛选较好的模型，减少人工操作。在理想情况下，可快速定位合适的模型。

本例应用 IBM SPSS Modeler 中的自动分类模型对算法进行粗略选择，如图 2.7 所示，对数据集进行初步分析，取得效果较好的 5 个分类模型，然后再基于分类模型独立分析，这样可以极大减少模型匹配和人工选择的时间，从而快速发现最佳的几个模型，有的放矢地进一步深入分析。



图 2.7 自动分类模型属性配置

在“模型”选项卡中设置模型数量为 5 个，在“专家”选项卡中选中所有的模型类型，可对模型参数进行修改，本例中使用默认的参数值。

运行后得到的自动分类模型筛选结果如图 2.8 所示，其中 CHAID 树、QUEST、CART、类神经网络、逻辑回归总体精确性较高。在模型列表中双击某一行的图形可查看因变量的分布情况，双击模型列表中的某一行可以查看模型的详情。

移动房车险数量

文件(F)　生成(G)　视图(V)　预览(P)

模型　图形　摘要　设置　注解

排序方式(S): 使用　◉升序　○降序　删除未使用模型　视图: 测试集

是否...	图形	模型	构建时间(分钟)	最大利润	最大利润发生比率	增益(前 ...	总体精确性 (%)	使用的字段编号	曲线下方面积
☑		CHAID 1	1	-32.436	0	1.856	94.303	12	0.699
☑		Quest 1	1	-48.734	0	1.000	94.303	36	0.5
☑		CART 1	1	-48.734	0	1.000	94.303	36	0.5
☑		类神经网络 1	1	-45.0	0	1.692	94.04	36	0.635
☑		Logistic 回归 1	1	-25.0	0	1.487	92.813	36	0.618

确定　取消　应用(A)　重置(R)

图 2.8　自动分类模型筛选结果

双击 QUEST 和 CART 模型时，发现其采用了返回固定值的方式来提高精确率，即将所有分类判断均返回 0 值(不投保移动房车险)作为结果，在不投保的样本比例较高时，这样操作当然也可以达到较高的精确率，但是没有实际的应用价值，所以在后续分析中将上述两种模型滤除。

在“图形”选项卡中可看出移动房车险投保比例较少，将鼠标悬浮在柱状图看到只有 348 条，占总数的 5.9%，如图 2.9 所示，其模型的分布情况与样本中目标变量的分布一致，说明模型本身没有实质的预测，即平均精确率全部来源于未投保样本的贡献(全部预测为未投保即可实现)。从预测变量的重要性中可以看出社会阶层、工作熟练程度、工作类别、教育程度、是否单身等重要性较高，但是这些变量之间的差异化并不明显。

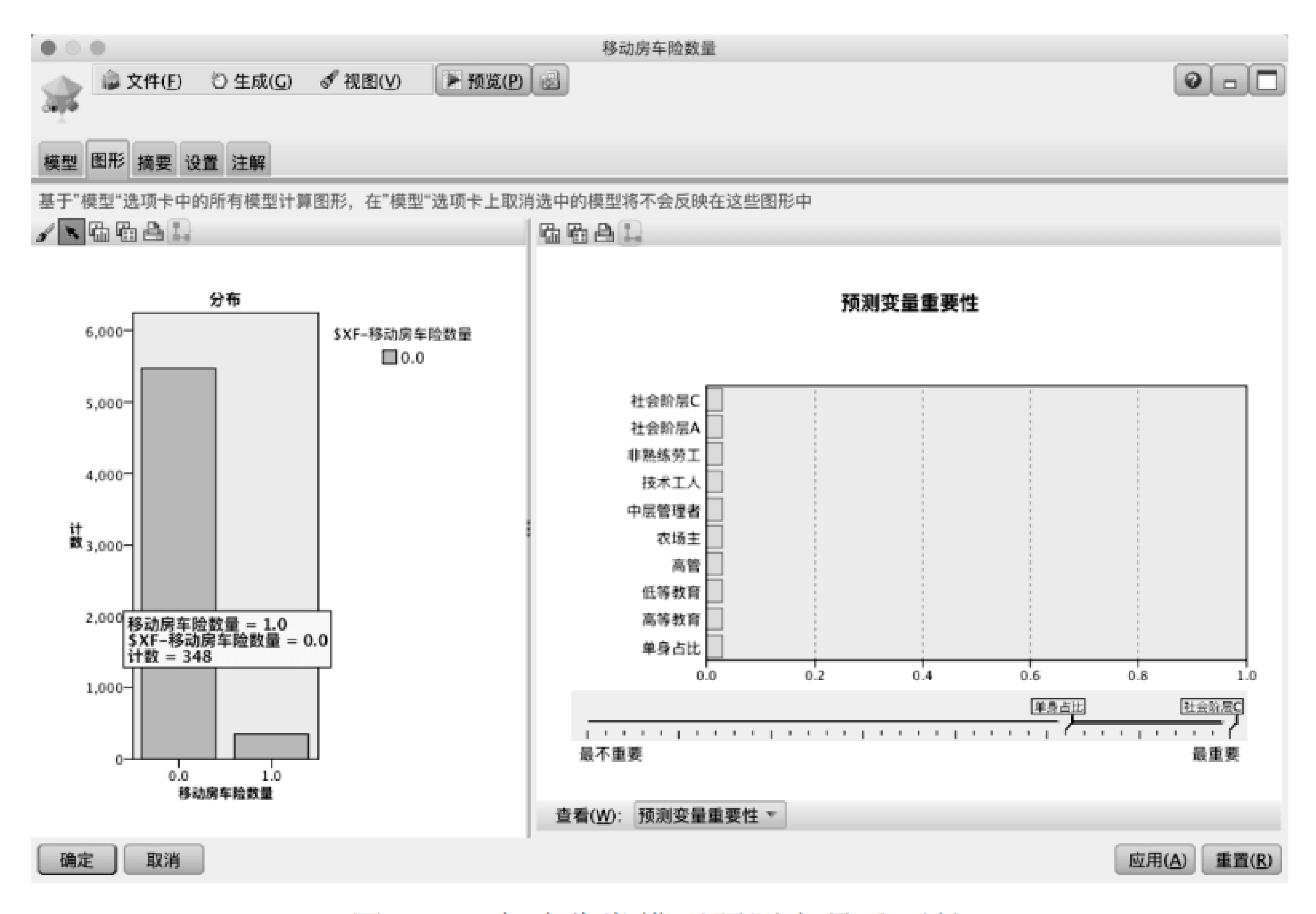


图 2.9　自动分类模型预测变量重要性

为了进一步验证自动分类器初选出的各个模型的效果，下面详细评估各模型的结果，并比较数据降维的效果。其中，为了比较降维对模型结果的影响，相应地选择 CHAID 树模型，比较降维前后的模型结果。同时，鉴于在逻辑回归模型中，不显著变量易对结果产生一定的影响，将逻辑回归的输入变量使用降维后的变量。

2.4.2 算法验证

依据模型的结果不仅可以看到各项指标是否达到要求，还要将结果与实际业务领域知识进行结合分析，以确认模型的应用价值，否则只注重单一的某几项结果指标，很可能出现过拟合的问题。

本例通过自动分类模型得出 CHAID 树的效果较好，所以单独分析 CHAID 树模型，其属性配置如图 2.10 所示，首先选择降维后的 36 个重要自变量作为输入，在“构建选项”中使用默认选项，即构建新模型且以构建单个树为目标，然后以降维前的 85 个变量作为自变量输入，比较降维前后对预测结果准确性的影响。





图 2.10　CHAID 树模型参数设置

运行模型后得到的结果如图 2.11 所示，展开所有层次可以看到模型中定义的分支全部为指向 0，即为投保移动房车险，说明模型并没有明确在何种情况下某一家庭会投保移动房车险，这种情况下模型的应用效果将不理想。

投保车险 = 4 or 投保车险 IS MISSING [模式: 0]
收入45-75 = 5 or 收入45-75 = 7 [模式: 0]
非熟练劳工 = 2 [模式: 0] ⇨ 0.0
非熟练劳工 IS MISSING [模式: 0] ⇨ 0.0
收入45-75 IS MISSING [模式: 0]
一辆车 = 0 or 一辆车 IS MISSING [模式: 0]
其他宗教比例 = 1 or 其他宗教比例 = 5 [模式: 0] ⇨ 0.0
其他宗教比例 IS MISSING [模式: 0]
企业家 = 0 or 企业家 = 4 [模式: 0] ⇨ 0.0
企业家 IS MISSING [模式: 0] ⇨ 0.0
一辆车 = 8 [模式: 0] ⇨ 0.0
投保车险 = 6 [模式: 0]
客户主类别 = 1 or 客户主类别 = 2 [模式: 0]
第三方私人险数量 = 0 [模式: 0] ⇨ 0.0
第三方私人险数量 = 1 [模式: 0] ⇨ 0.0
客户主类别 = 3 or 客户主类别 = 8 or 客户主类别 = 9 [模式: 0]
第三方私人险数量 = 0 or 第三方私人险数量 IS MISSING [模式: 0] ⇨ 0.0
第三方私人险数量 = 1 [模式: 0] ⇨ 0.0
客户主类别 = 4 or 客户主类别 = 5 or 客户主类别 = 6 or 客户主类别 = 7 or 客户主类别 = 10 [模式: 0]
购买力水平 = 3 [模式: 0] ⇨ 0.0
购买力水平 IS MISSING [模式: 0] ⇨ 0.0

图 2.11　CHAID 树模型结果

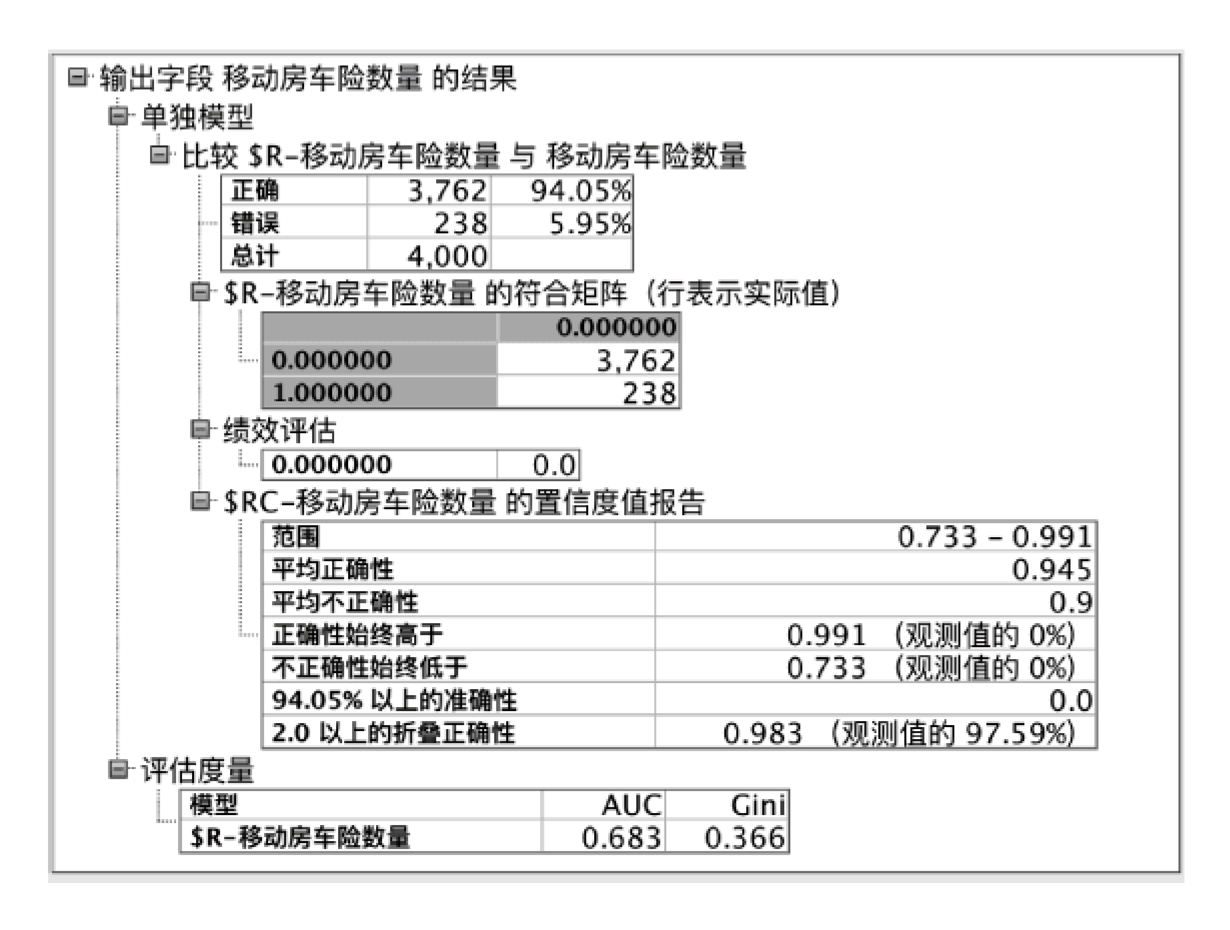
运行模型后使用测试数据作为模型的验证数据，用它来替换模型训练过程中的样本输入，即将 eval.xlsx 输入连接到生成的模型中，并在模型后连接“分析”节点，选中“重合矩阵”“绩效评估”“评估度量”“置信度图”作为输出指标，如图 2.12 所示。



图 2.12　CHAID 树模型分析参数配置

运行分析节点得到分类评价指标的结果，如图 2.13 所示，可以看到其正确率达到 94.05%，AUC 和 Gini 系数这两个指标的结果也较好。此外，使用 36 维自变量的指标与之

相比仅仅在 2.0 以上的折叠准确性上略有差别，为 0.981，说明变量降维后对 CHAID 模型的准确性结果影响不大，降维后的结果具有一定的代表性。



输出字段 移动房车险数量 的结果

单独模型

比较 $R-移动房车险数量 与 移动房车险数量

正确	3,762	94.05%
错误	238	5.95%
总计	4,000	

$R-移动房车险数量 的符合矩阵（行表示实际值）

	0.000000
0.000000	3,762
1.000000	238

绩效评估

0.000000	0.0

$RC-移动房车险数量 的置信度值报告

范围	0.733 - 0.991
平均正确性	0.945
平均不正确性	0.9
正确性始终高于	0.991 （观测值的 0%）
不正确性始终低于	0.733 （观测值的 0%）
94.05% 以上的准确性	0.0
2.0 以上的折叠正确性	0.983 （观测值的 97.59%）

评估度量

模型	AUC	Gini
$R-移动房车险数量	0.683	0.366

图 2.13 CHAID 树模型分类效果评估

如果仅看模型的准确率指标，模型的分类效果很好，但是我们发现 CHAID 树与 QUEST、CART 的问题类似，在分类过程中主要是返回 0 值样本的值，其分类的准确性与样本中因变量为 0 值的比例是一致的，观察模型的评估结果可以发现移动房车险数量的符合矩阵中并没有出现值为 1 时的预测值，说明在预测过程中模型对于投保这一险种的预测结果全部为 0，即不投保移动房车险，明显不符合实际业务，所以无论是否使用降维操作，这个模型也不具有应用价值。

片面看重模型中的准确性等直接指标往往易被其欺骗，即使通过了样本的测试，在实际应用中也难以有效应用，这个家庭会投保移动房车险吗？系统并不需要任何运算，只要回答"不会"就会达到 94%的准确率，但是这样的模型明显没有实际意义。

将自动算法选择的 CHAID 模型淘汰后，继续分析验证逻辑回归模型，对降维后的样本应用逻辑回归模型，移动房车险投保作为因变量，由于逻辑回归模型中不显著，自变量对模型的准确率影响较小，将其排除，使用降维后的 36 维字段作为自变量。与 CHAID 模型类似，逻辑回归模型的整体准确率较高，但是召回率极低，模型出现了过拟合的情况，可以预见其在实际应用中效果不好，所以逻辑回归模型也无法直接在保险公司中实际使用。

下面接着分析验证类神经网络算法的结果，模型准确率为 93.05%，具有较好的分类效果，但是与逻辑回归类似，其 AUC 和 Gini 系数指标表现一般。其中召回率为 0，说明直接应用模型时基本上没有实际意义。

综上，上述几种模型均未通过验证，比较几种模型未通过验证的原因，发现在分析过程中样本集存在严重的不平衡问题，导致算法"投机取巧"几乎不做分析，直接返回多样本记录对应的结果值，使各种分类算法在实际应用中均失效，虽然具有较高的整体准确性，却无法应用于实际企业运营中，这种情况在现实生活中非常多，如疾病预测、流失客户预警、欺诈检测、垃圾电话或邮件检测等，处理不平衡数据集是目前数据分析领域比较热门的问题之一，对其进行评价的指标主要是以 ROC 曲线来进行评价，而 ROC 曲线没有给出具体的参数值，所以采用 ROC 曲线下方的面积来评价，即采用 AUC 指标。

2.4.3　算法优化

不同模型选择和验证为常规选择方法，大多数业务均可进行上述操作，但是数据集的特点不一，且分析任务的准确性要求具有个性化，需要对模型进一步优化，以达到实际应用的需要。模型优化的方法主要从业务和建模技术等思路上进行优化，前者要对业务和数据特征有深入的理解，后者则主要靠的是挖掘技巧。

从前面的分析中可以看出，上述算法均不能直接应用于实际的业务中，需要对算法选择过程进行优化。由于样本存在不平衡，所以并非优化算法的准确率，而是 AUC 这一指标，也可应用召回率对模型结果进行比较。

对分析流程重新优化后的结果如图 2.14 所示，对逻辑回归和 CHAID 进行了模型参数改进。经过分析，发现数据降维使模型降低了准确性，所以在新的流程中不再使用“特征选择”，直接将分区数据应用到模型中。由于 SPSS Modeler 中的“平衡”节点每次运行都会随机产生新的样本记录，结果并不固定，容易对模型结果产生干扰，所以平衡后的样本集导出到 Excel 文件，然后再从 Excel 中读取出来，只要运行一次，就生成一个平衡后的数据源，此数据源(balance. xlsx)作为新流程的输入，样本数据不再变化。

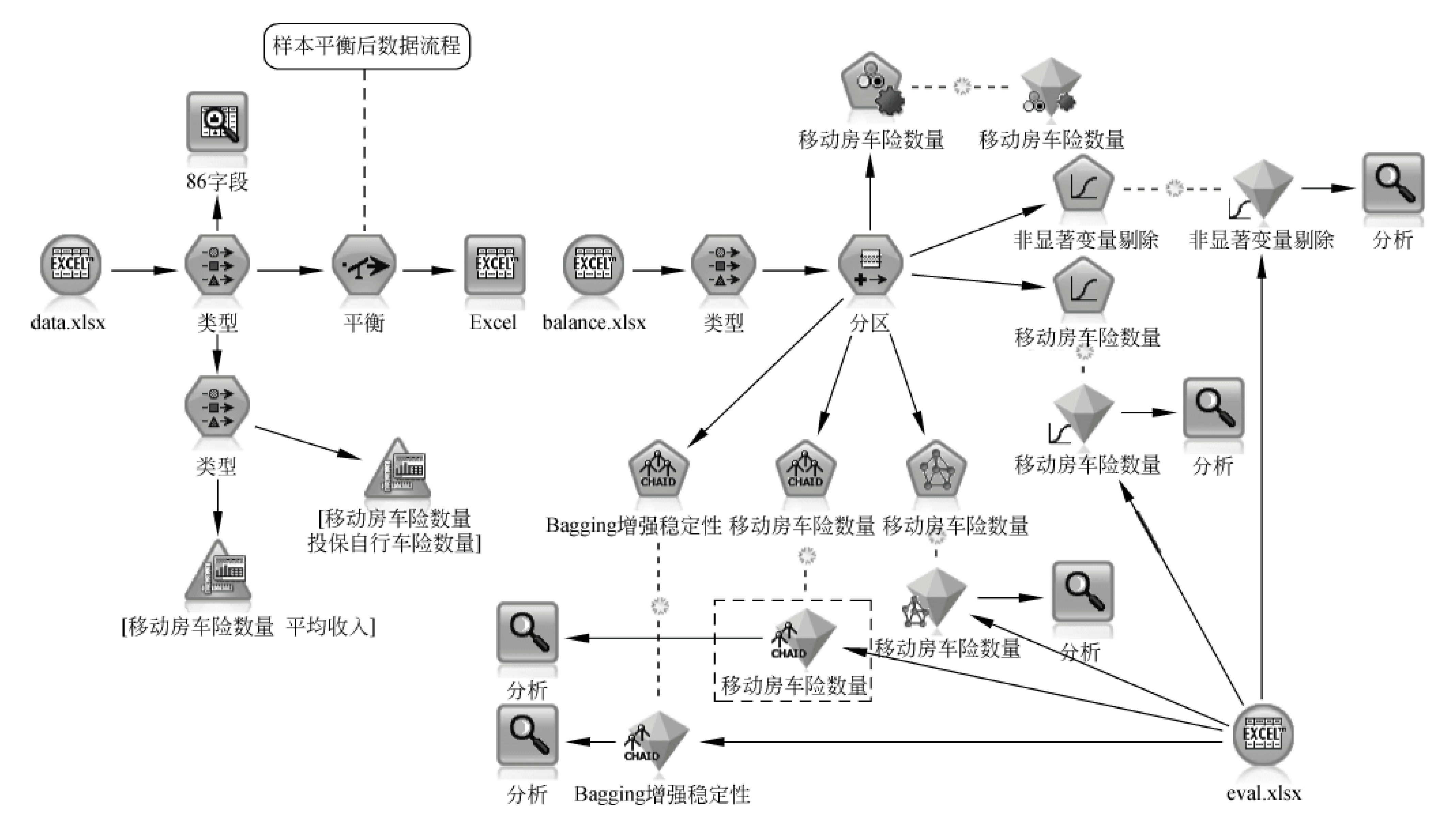


图 2.14　平衡节点参数配置

2.4.4　平衡数据集

现实中目标变量的样本分布经常出现已经失真的情况，至少有一个类别的样本数量占比小于 20%，实际上很多类别的占比甚至低于 5%，如信用卡诈骗、疾病检测等，而数据挖掘却要从中找出小概率的样本来，要实现可靠的数据结果，需要根据样本数据的特点正确平衡数据，平衡之后有助于建立能够真正解决业务问题的模型。另外，数据平衡之后还要监控其变化，因为随着时间的推移，样本的结构可以发生变化，需要定期进行调整，以保持模型的准

确性。

数据挖掘中的“平衡”类似于统计学中的加权，为了使分类中的样本比例更加合理，常用的样本平衡包括基于采样的方法、对正负样本采用不同的代价函数、使用集成的方法等，假如正样本数量远远小于负样本数量，那么可以采用这些方法进行平衡。

基于采样的方式，对正样本过采样或对负样本进行欠采样，也可混合方式，即增加正样本的同时减少负例数量，这种方式最简单，只是调整样本集就可以达到平衡。由于过采样是通过复制的方式实现的，其不足之处是会使变量的方差比实际的小。当然，过采样不会丢失样本，包括样本的误差。而欠采样一般采用随机删除样本的方式，容易去掉样本的重要特征，且其方差比实际值要高。R 语言和 Scikit-learn 均有相应的工具方法可以用于调整样本的权重值。在采样方式选择中，如果正负样本数均较少，使用过采样，如果正负样本数均足够多，采用下采样或混合方式。

在模型损失函数中调整惩罚项的权重，增加正样本的权重，减少负样本权重，这种方法的难点要根据实际情况来设置权重值，在实际应用中一般让各个分类的加权损失值近似相等，但是由于矩阵是稀疏的，这一方法在实际应用中效果不是很好。

集成的方式，应用 SMOTE、Boosting、Bagging 等。SMOTE 是 Synthetic Minority Oversampling Technique 的缩写，即合成少数类过采样技术，其思想是在正例中创造出一些新的样本，其缺点是只能生成某一范围内的样本，不能创造超出少数类样本之外的新样本，且容易增加类间重叠的可能性。而 Boosting 和 Bagging 算法主要是集成多个弱分类器得到更合理的边界，实现更好的分类效果，Boosting 更关注被错分的样本。需注意的是，Boosting 的重采样并非是样本，而是样本的分布，经过迭代使被分错的样本逐渐划分到下一次的训练集，优点是简单，也不用担心出现过拟合，缺点是噪声点和异常点敏感，因为每天迭代噪声样本的权重都会被放大，由于其迭代时无法并行计算，所以运行速度较慢。

统一分类的方法，将少数类样本统一成多数类中，将分类划分问题转化为异常检测，这种方法重点不在于捕捉类间的差别，而是为其中的一类进行建模，然后将少数类作为异常样本进行检测，采用这一方法的前提是训练集中数据质量较高，如果其本身含有较多噪声数据，容易产生较大误差。

上述几种方法在实际应用中效果评价很关键，评估时无论采有何种方法进行训练，最终要使用实际分布的测试集进行检验。另外，不能使用准确率这一指标，应该使用 ROC 曲线、准确度召回曲线、利润收益曲线等方式对结果进行可视化，或者使用 AUC、召回率、F1 分值等指标定量评估，不要迷信分值，而要专注于少数类的分类正确率。

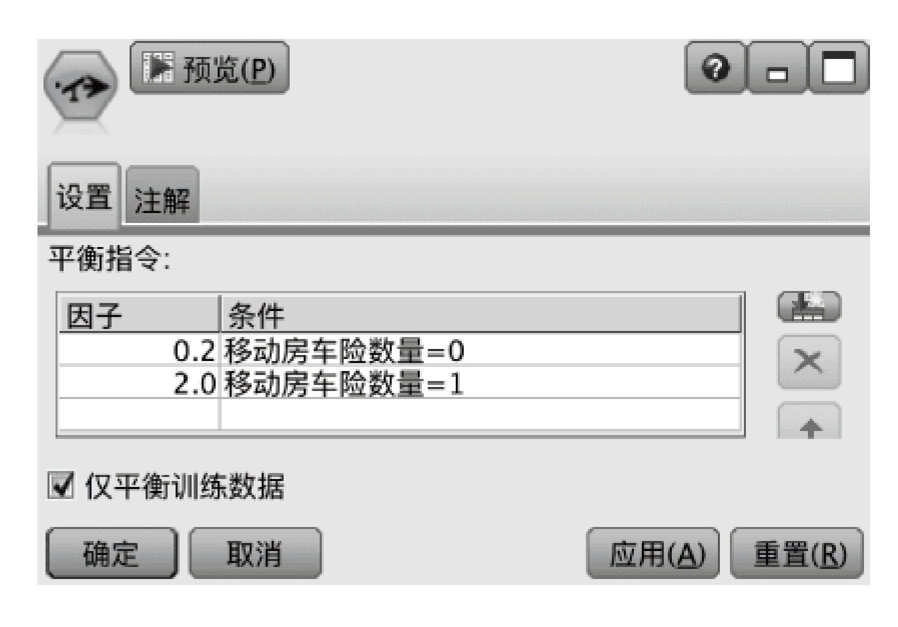


图 2.15　平衡节点参数配置

考虑到样本数据中未投保移动房车险的记录占比为 94%，所以在之前的模型选择中其结果虽然准确率较高，但是 AUC 值和召回率的分值并不高，在 IBM SPSS Modeler 中的可使用“平衡”节点对样本记录进行随机采样，如图 2.15 所示，将未投保移动房车险的记录数降为原来的 20%，大约为总记录数的 18%，而投保记录数量并不一定需要与未投保样本数据平均分布，所以只升为原来的 2 倍，其记录数量与未投保数量大体一致，约为

12%。因子设置过程中要依据实际业务情况来设置，没有固定规则，但是要避免过多地复制少数类样本，或对多数类样本过度欠采样。

经过平衡后，再次使用自动分类模型进行算法探索，但在算法排序规则中选择 AUC 作为评价指标，即曲线下的面积越大，其模型排名越靠前，结果如图 2.16 所示，前 5 个模型依次为逻辑回归、类神经网络、CHAID、QUEST、CART，虽然以曲线下的面积作为评价标准，但各模型的总体精确性也呈现依次降低的顺序，说明各模型未被样本数据“欺骗”，从“图形”选项卡中的结果也可以看出，对投保移动房车险的预测总体精确性高于 60%，AUC 值最高为 0.779，具有一定的应用价值。

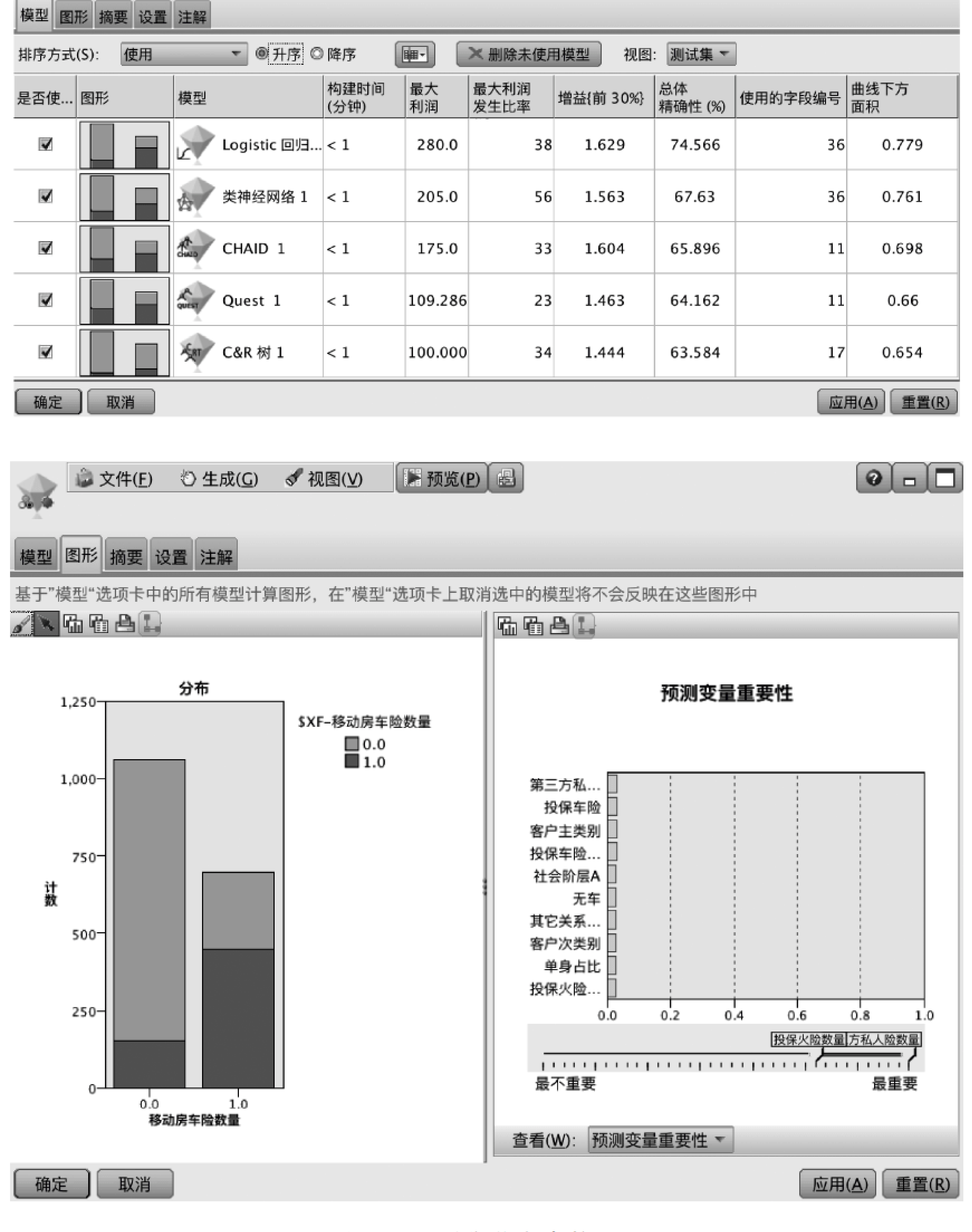
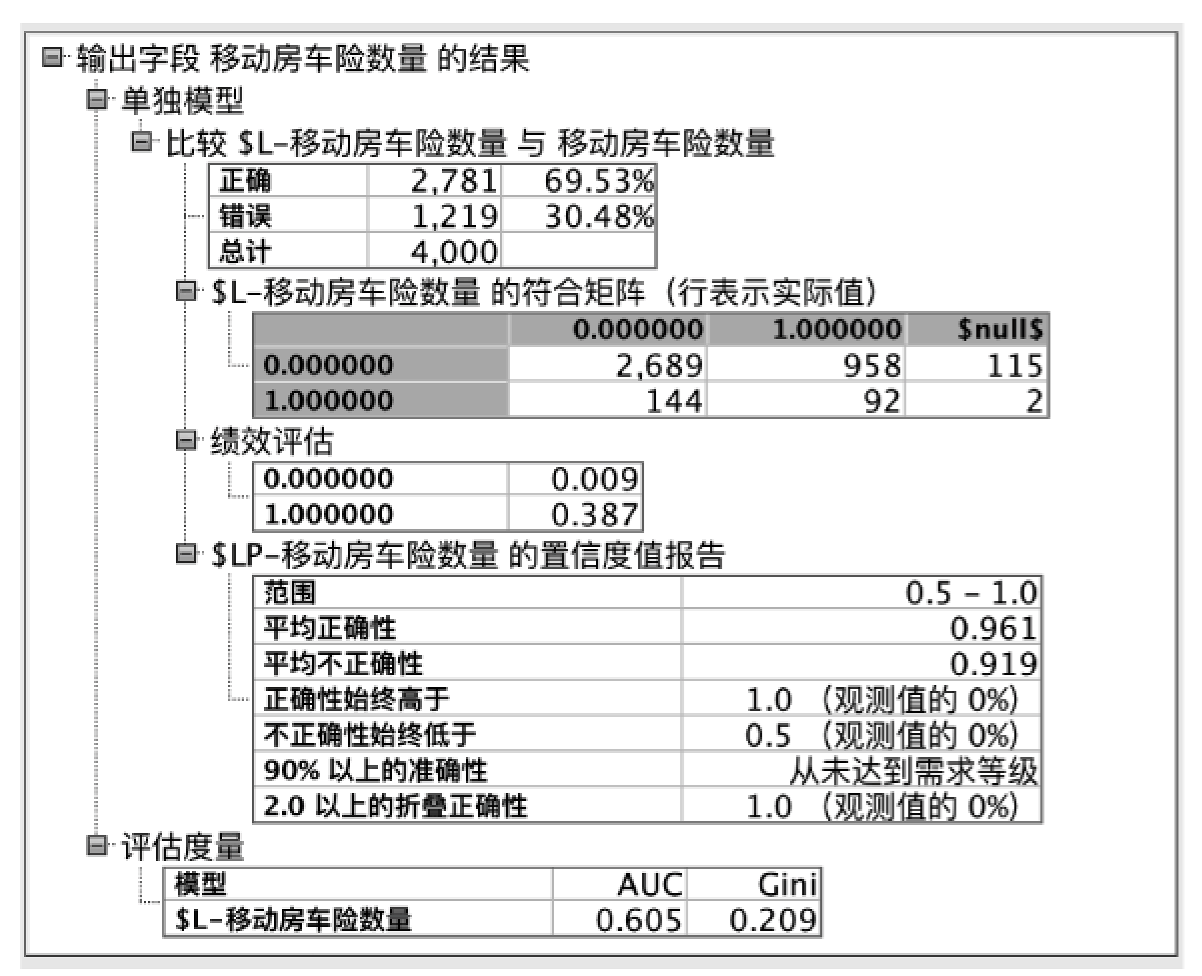


图 2.16　平衡节点参数配置

经过逻辑回归模型运算，得到如图 2.17(a)所示的分析结果，AUC 值较低，Gini 值也较低，可以看到模型的分类能力提高了，模型的召回率和 AUC 指标值也提高了，说明模型的应用价值有所提高。



输出字段 移动房车险数量 的结果

单独模型

比较 $L-移动房车险数量 与 移动房车险数量

正确	2,781	69.53%
错误	1,219	30.48%
总计	4,000	

$L-移动房车险数量 的符合矩阵 (行表示实际值)

	0.000000	1.000000	$null$
0.000000	2,689	958	115
1.000000	144	92	2

绩效评估

0.000000	0.009
1.000000	0.387

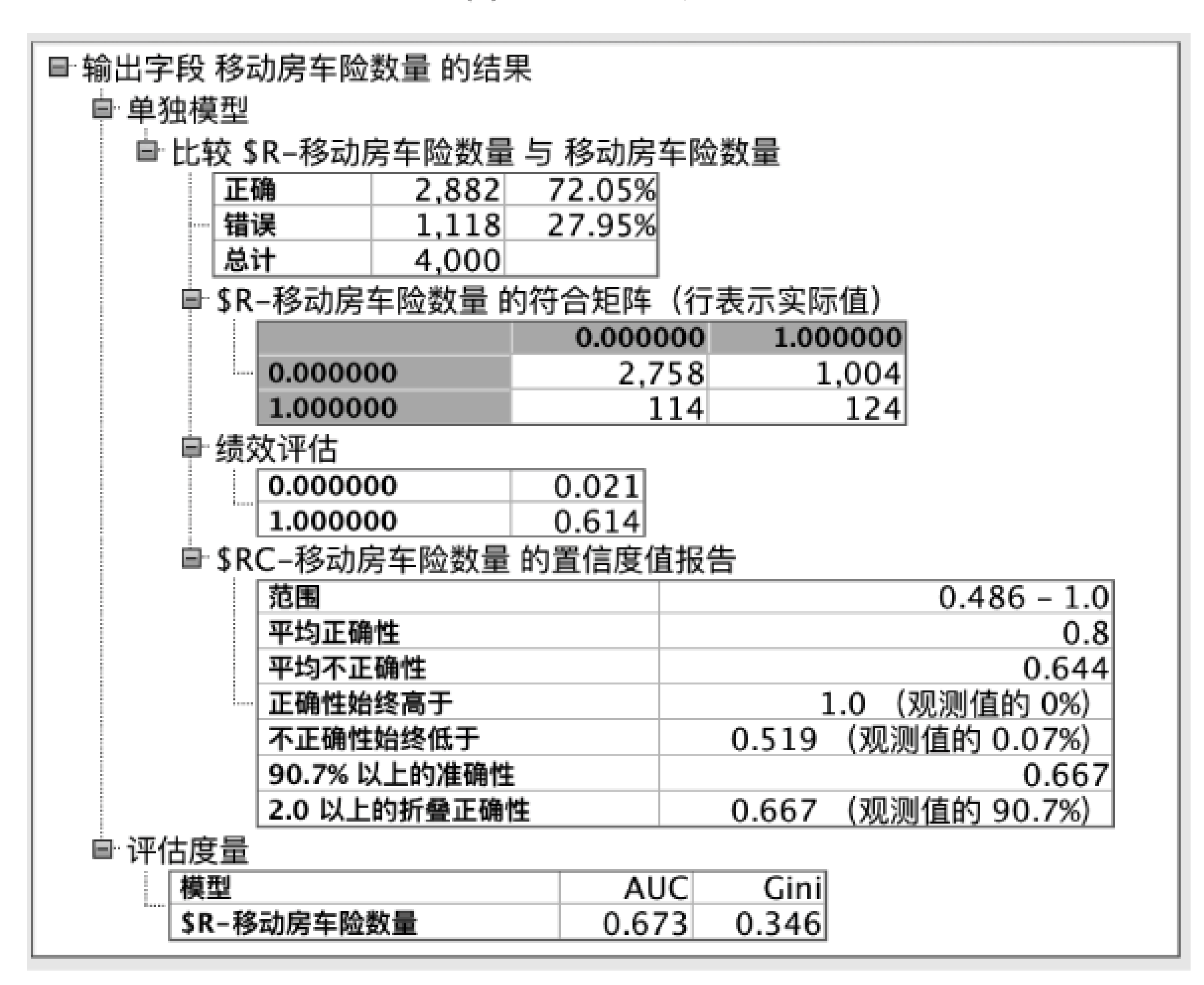
$LP-移动房车险数量 的置信度值报告

范围	0.5 - 1.0
平均正确性	0.961
平均不正确性	0.919
正确性始终高于	1.0 (观测值的 0%)
不正确性始终低于	0.5 (观测值的 0%)
90% 以上的准确性	从未达到需求等级
2.0 以上的折叠正确性	1.0 (观测值的 0%)

评估度量

模型	AUC	Gini
$L-移动房车险数量	0.605	0.209

(a) 逻辑回归模型



输出字段 移动房车险数量 的结果

单独模型

比较 $R-移动房车险数量 与 移动房车险数量

正确	2,882	72.05%
错误	1,118	27.95%
总计	4,000	

$R-移动房车险数量 的符合矩阵 (行表示实际值)

	0.000000	1.000000
0.000000	2,758	1,004
1.000000	114	124

绩效评估

0.000000	0.021
1.000000	0.614

$RC-移动房车险数量 的置信度值报告

范围	0.486 - 1.0
平均正确性	0.8
平均不正确性	0.644
正确性始终高于	1.0 (观测值的 0%)
不正确性始终低于	0.519 (观测值的 0.07%)
90.7% 以上的准确性	0.667
2.0 以上的折叠正确性	0.667 (观测值的 90.7%)

评估度量

模型	AUC	Gini
$R-移动房车险数量	0.673	0.346

(b) CHAID模型结果

图 2.17 分类平衡后的模型结果

CHAID 的模型正确率为 72.05%，AUC 和 Gini 系数分别为 0.673 和 0.346，相较于逻辑回归均有一定提高，从移动房车险的符合矩阵中也可以看到对于投保的预测正确比例要超过预测错误比例，说明 CHAID 模型更具有应用价值。

2.4.5 修改模型参数

模型参数在一定程度上依赖于样本的特点和模型的原理，在决策树模型中遇到不平衡样本可以使用增强稳定性 Boosting 或 Bagging 技术进行改进，像逻辑回归算法可使用前进法或后退法逐步剔除非显著性变量来改进模型，而支持向量机(SVM)模型中可以修改惩罚

参数 L2 等来调优。大多数情况下，参数调优的过程是反复渐进式的，很难一蹴而就。

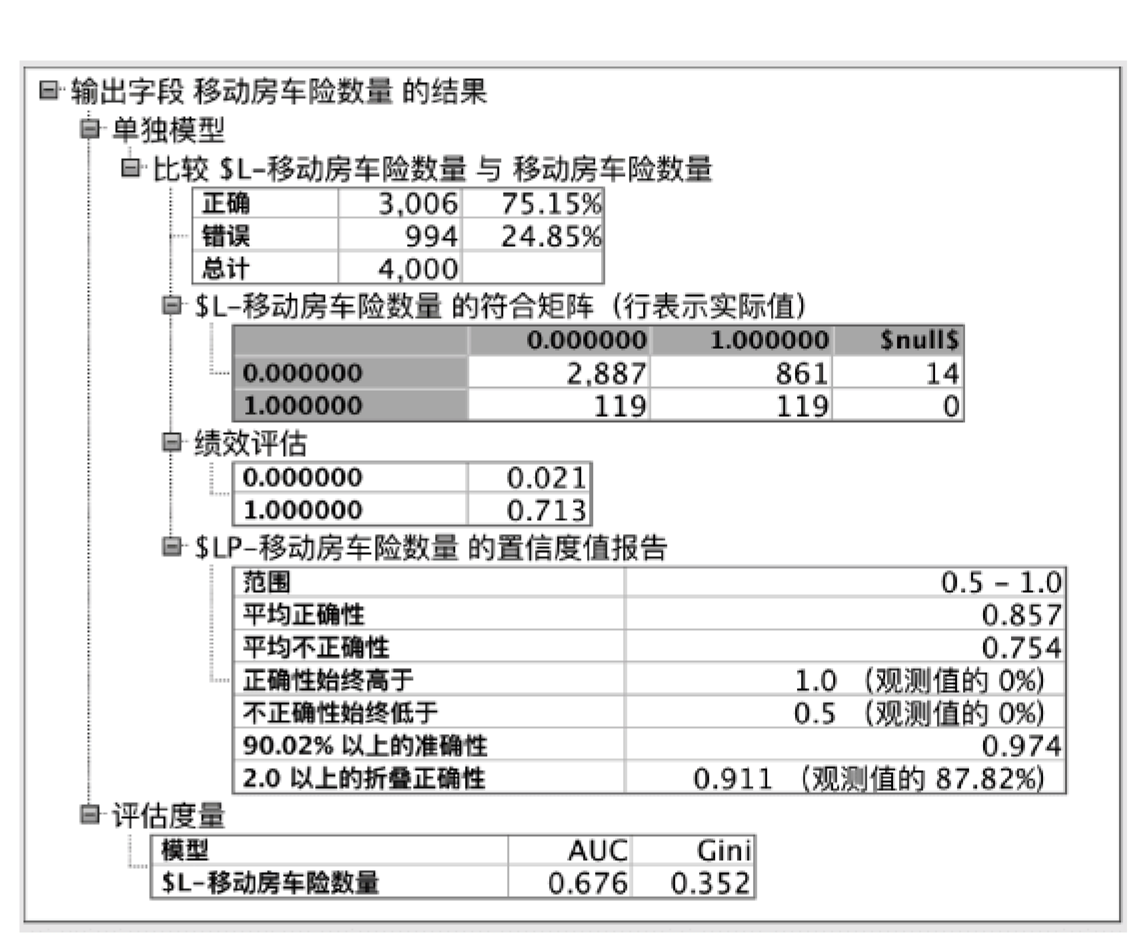
1. 逻辑回归模型参数调优

为了进一步提高算法的 AUC 等指标，通过改进模型参数来继续优化模型，在逻辑回归模型中可选择“多项式”或“二项式”，数据集的目标变量有多个分类时选择前者，只有两种类别时选择后者，本例中理论上可选二项式，但是模型结果改进效果有限。

比较选择“进入法”“前进法”“步进法”等方法后的模型结果，看是否改进，其中进入法并不对输入变量做选择，全部用于模型，前进法或步进法则是先生成一个简单模型，然后逐渐增加输入变量，直到新加入的变量不再提高模型的准确性为止，向后法则相反，先由所有输入变量生成模型，然后逐渐移除对模型影响最小的输入变量，直到无法删除输入变量为止，从而生成模型，本例中使用前进法来优化模型。

除此之外，通过对模型的输入变量手工过滤来提高性能，先对变量进行一次逻辑回归，然后将其中不显著的自变量($p>0.05$)剔除，重新进行逻辑回归，这样可以使模型的准确率和 AUC 均有明显提升，但剔除较多变量后模型的伪 R 方值略有下降。

本例对“房产数”“同居占比”等自变量进行剔除后得到如图 2.18 所示的结果，可以看到，准确率从未调整参数前的 69.53%提高到 75.15%，且 AUC 的值从 0.605 提高到 0.676，改进程度较大。



输出字段 移动房车险数量 的结果

单独模型

比较 $L-移动房车险数量 与 移动房车险数量

正确	3,006	75.15%
错误	994	24.85%
总计	4,000	

$L-移动房车险数量 的符合矩阵（行表示实际值）

	0.000000	1.000000	$null$
0.000000	2,887	861	14
1.000000	119	119	0

绩效评估

0.000000	0.021
1.000000	0.713

$LP-移动房车险数量 的置信度值报告

范围	0.5 – 1.0
平均正确性	0.857
平均不正确性	0.754
正确性始终高于	1.0 （观测值的 0%）
不正确性始终低于	0.5 （观测值的 0%）
90.02% 以上的准确性	0.974
2.0 以上的折叠正确性	0.911 （观测值的 87.82%）

评估度量

模型	AUC	Gini
$L-移动房车险数量	0.676	0.352

图 2.18　逻辑回归模型参数改进后的结果

2. CHAID 模型参数调优

Bagging 和 Boosting 都是用来提高模型准确率的方法，其中 Bagging 是 Bootstrap Aggregating 的一种，根据均匀概率分布从数据集中有放回地抽样提取样本，形成多个训练集，然后得到多个训练集预测结果，并对结果采用投票的方式处理分类问题，采用平均的方法处理回归问题。Boosting 是通过不断调整训练失败的样本较大权重，即根据错误率来取样，通常比 Bagging 方法具有更高的准确率，当然，也有可能会引起过拟合。

在 SPSS Modeler 中可以使用这些技术的模型有神经网络、CHAID、QUEST、CART、线性回归等，在对数据特征未深入理解的情况下，可分别测试并比较 Boosting 和 Bagging 两种选项的结果，选择结果较优的选项。

本例对 CHAID 模型使用 Bagging 技术进行优化，并选择“似然比”作为类别目标的卡方，成分模型的数量设为 10，得到较优的结果如图 2.19 所示，可以看到模型的准确率提高到 78.77%，AUC 和 Gini 系数指标分别为 0.691 和 0.383，综合来看，模型的改进程度较大。

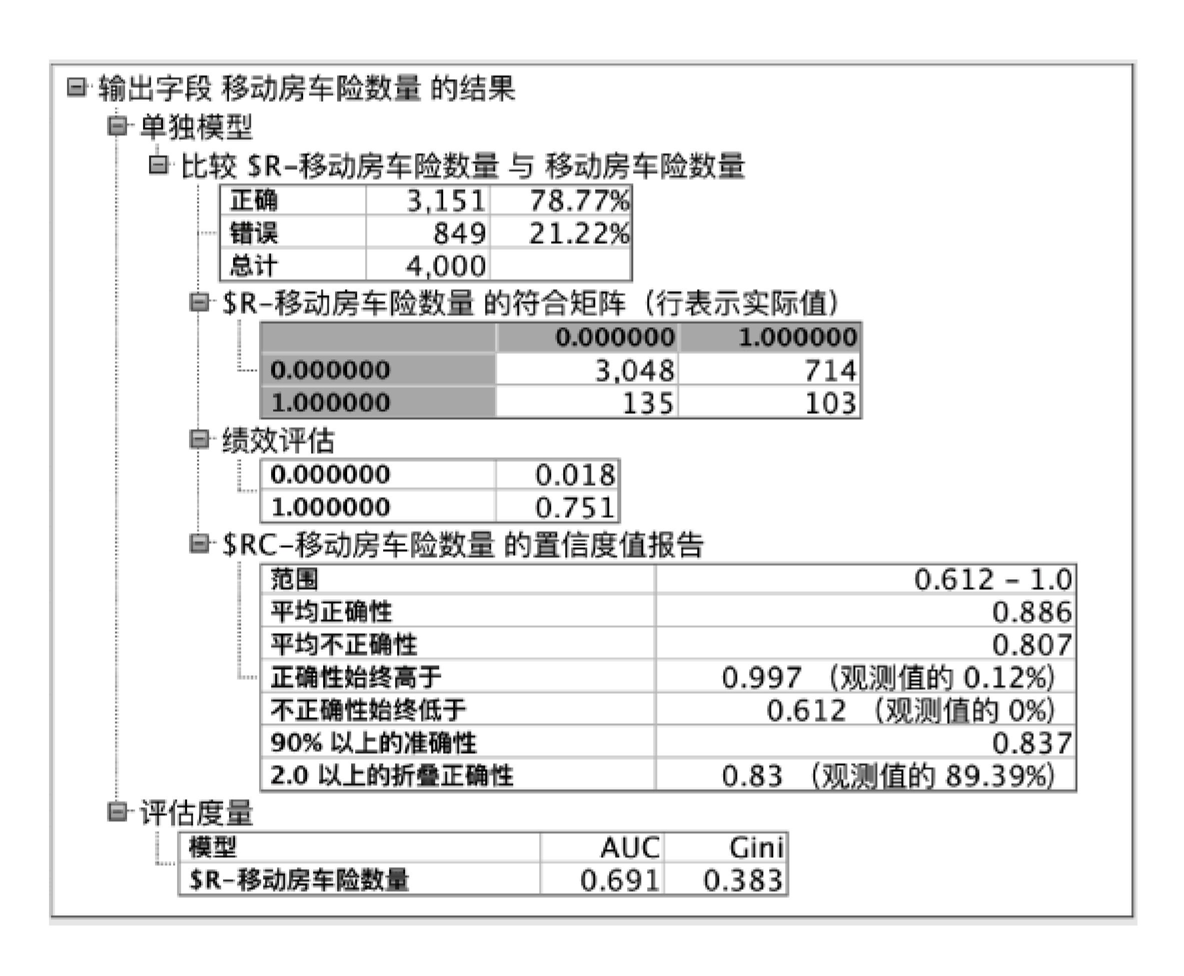


图 2.19 CHAID 模型参数改进后的结果

需要注意的是，应用 Bagging 或 Boosting 选项后，模型的计算量会大幅增加，是原来的 N 倍（成分模型的数量，默认为 10）。所以，训练集样本数较多时，训练时间明显变长，特别是 Boosting 方法，因为它的预测函数是串行生成的，不支持并行训练。

2.5 总结

数据探索是模型选择中的一个重要环节，首先要分析目标变量的特点，决定模型的选择范围，然后分析自变量中数据分布的特点，通过可视化技术将分布情况以图形化方式展示，不仅可以对样本有直观的认识，也可以大致推断出其与目标变量之间的关系。模型结果出来后，还可以用于验证结果的合理性。在了解数据特点的基础上选择模型可以减少很多工作量。在本例中，如果一开始就清楚目标变量中具有很严重的不平衡问题，就可以直接在模型选择前先对其做平衡处理。

处理样本平衡时要与业务目标结合，检查模型的混淆矩阵，看各分类中的比例是否与业务要求一致，像欺诈这种小概率事件检测中，要注意模型样本数很少的分类易被模型忽略，导致预测准确率虚高，这类业务下平衡数据集并不需要将分类数据均匀分布，通过采样方式减少某一分类样本记录数时，要避免过度采样导致关键样本特征丢失，也要注意不可过多复制单条样本，防止人为放大某一数据特征，对上述情况的预防措施就是使用独立的测试集对模型进行检测，以确定模型真实有效。此外，需要注意某些分析任务中要求目标变量不同类样本平衡具有固定比例时，如性别比要求 1∶1 等，而企业运营中数据通常是动态的，平衡系数也需要定期手动调节。

模型选择中不要盲目相信机器自动化的选择，由于机器并不熟悉业务，其对于模型的评

价指标无法与业务规则相对应，容易导致虚假的高性能模型结果。从本例中还可以看到对模型评价指标解读的重要性，如果不能从评价指标中发现问题，直接应用模型到业务系统中将无法带来有益的作用，发现模型问题的能力也是数据挖掘人员的一项重要实践技能，发现问题后替换模型或对模型不断地调整参数，使其结果逐渐逼近业务目标要求，最终才可能在业务中应用。

逻辑回归模型中的预测或分类是通过回归方程实现的，观察逻辑回归方程的系数，其值为正则说明具有正向影响，在本例中可以看到高管、社会阶层 A、社会阶层 B、平均收入、投保车险、投保房车险、投保火险等具有较高的正系数值，这与描述性统计结果一致，即高收入中产以上阶层且对家庭中重要资产投过保的用户是移动房车险的重要目标客户群体。相反，社会阶层 D、投保寿险、投保身残险的用户基本不会购买此险种。

第3章

常用可视化的多维分析

在数据分析中，通过各种可视化的图形，从多个维度、多个层次展示企业商务的执行情况，发现可能存在的问题或潜在的危机，并预测未来业务发展的趋势，具有重要的价值。而且利用可视化的工具，也可以发现数据的一些质量问题、分布特点，可以为进一步的数据挖掘做预处理。因此，结合业务理解，利用常用的可视化工具，如 Tableau、Lumira、国内的永洪大数据分析工具 Yonghong Z-Suite 等，对数据做一定深度的分析，这是数据分析师的基本功。

可视化图形通过位置、长、宽、角度、大小、色调、形状等多个方面，以视觉效果来表达图形相关含义。在数据分析中，各种不同的图形具有各异的作用，这也就为解决多元问题，深层次了解业务逻辑提供了方法途径。

可视化图形的作用各异，箱图的作用是展现数据的离散状态，以其数据节点：上限、下限、上四分位、下四分位、中位数及异常值为依据，来分析数据的离散程度等信息，可应用于数据预处理，识别数据异常值及分析数据离散状态。雷达图的作用是对事物的不同维度进行分析研究(通常，维度应大于或等于四维)，通过网状结构的图形对比形象展示各维度属性的相关状态。标签云的作用是显示词频，将标签出现或者被引用的多少，通过标签字体的大小和颜色等视觉效果呈现出来。气泡图的作用是研究数据之间的关系，以气泡的位置和大小及颜色来表现变量之间的关联。树图的作用是展现数据的层次关系，通过树图区域模块的占比、颜色深浅及层次等信息来研究数据之间的逻辑结构关系。地图的作用是展示数据与地理位置之间的关系，同时，可以根据颜色的深浅来判断地理区域或关键词的热门程度。高低图的作用是展现数据的波动特性，其不仅能研究数据长期波动的特性，也能研究数据短期的波动特性。双轴图的作用是展现数据的波动特征以及其数据之间的关联，通过在同一分析图形中绘制不同类别的图形，形象地展示数据之间的关系。关系图的作用是展现事物之间的相关性及其逻辑结构，以事物之间连线的粗细和颜色深浅等视觉效果为依据，研究事物之间复杂的逻辑关系。热图的作用是表现数据的热点特征，以视觉化的区域和色彩来表现数据的热点程度等特征。

在种类繁多的分析图形中，需要根据研究问题的不同，选取适当的分析图形来进行数据分析。下面对常用的分析图形作简要介绍。

3.1　箱图

箱图是一种显示数据离散状态的分析图形。通过箱图，能够获得相关数据节点等信息。箱图主要包含 6 个数据节点：上限、下限、上四分位、中位数、下四分位和异常值。

通常来说，上限位于上四分位加上 1.5 倍四分位距处，下限位于下四分位减去 1.5 倍四分位距处（如果样本数据在上四分位加上 1.5 倍四分位距及下四分位减去 1.5 倍四分位距处无数据，则上下限即为样本数据的最大值和最小值，这种情况下就没有异常值了）。在箱图中，异常值定义为数据点在样本数据中的位置大于上四分位加上 1.5 倍四分位距或小于下四分位减去 1.5 倍四分位距的数据。其中，温和异常值使用“o”表示，极端异常值使用“ * ”表示。

在箱图中，箱子占据了样本数据的一半，因而，箱子的宽度在一定程度上反映了数据的波动程度。箱子中间的一条线代表了中位数，其反映了样本数据的平均水平，同时，当中位数偏离上四分位和下四分位中心位置时，数据就表现出一种偏态性，中位数越偏离箱子中心位置，偏态性越强。箱图的另一主要功能是识别数据异常。进行数据分析时，异常数据可能会对分析结果造成影响，因而，通过箱图识别出异常值，并将其剔除，这将有利于数据分析结果的正确性。与其他统计图形相比，箱图可以将多批数据放在同一坐标轴上，并排排列进行对比，使得样本数据特征的分析变得更加容易。

为了更加形象地了解箱图相关结构特点及功能，结合香水实例，使用 SPSS Statistics 工具绘制出的箱图如图 3.1 所示，统计分析香水价格的相关情况。

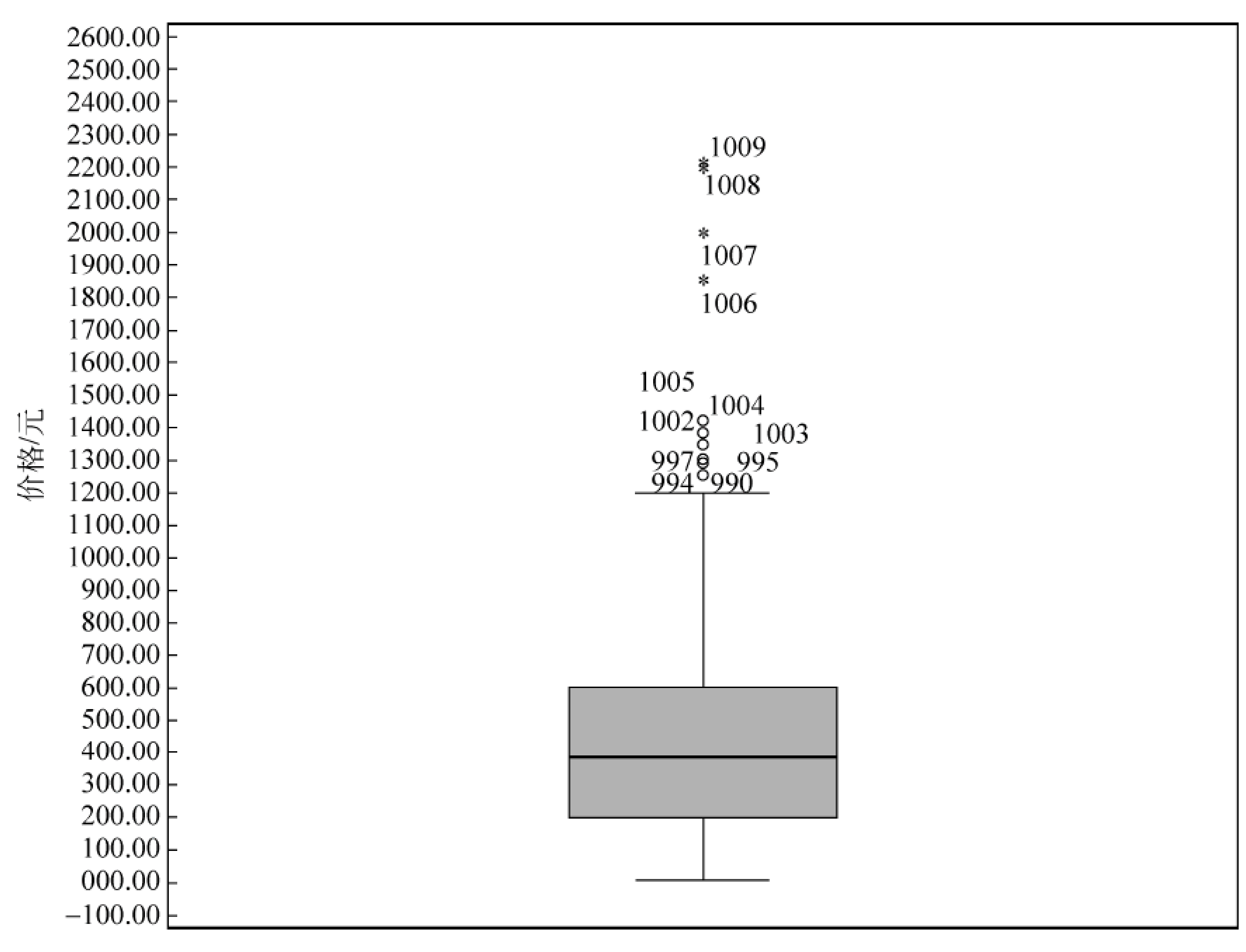


图 3.1　香水价格的箱图

图 3.1 所示箱图有关的几个数据如下。

下限：9.9，由箱子下方的一条线表示，下限由下四分位数减去 1.5 倍四分位距决定。下四分位：200，由箱子的下边线表示，代表数据的下四分位。中位数：385，由箱子中间的一条线表示，代表数据的中位数，反映了香水价格的平均水平为 385。上四分位：600，由箱子的上边线表示，代表数据的上四分位。上限：1189，由箱子上方的一条线表示，上限由上四分位数加上 1.5 倍的四分位距决定。

从图 3.1 中可以看到大于上限的圆圈点，这些点就是异常值，分析数据时可将其忽略。此外，这些数据点对应的标号是这些异常点在样本数据之中的位置，可以根据这个位置信息找到该异常点在原始数据中的具体位置。994～1005 号数据都是温和异常值，用“o”来表示；1006～1009 号数据都是极端异常值，用“＊”表示。

绘制箱图前，有可能需要对数据进行预处理。举例说明：针对香水样本数据，评价量在一定程度上反映销售量，对“探究不同品牌香水评价量相关特征”这一问题进行分析，在未对数据进行预处理之前，使用 SPSS Statistics 工具绘制箱图，如图 3.2 所示。

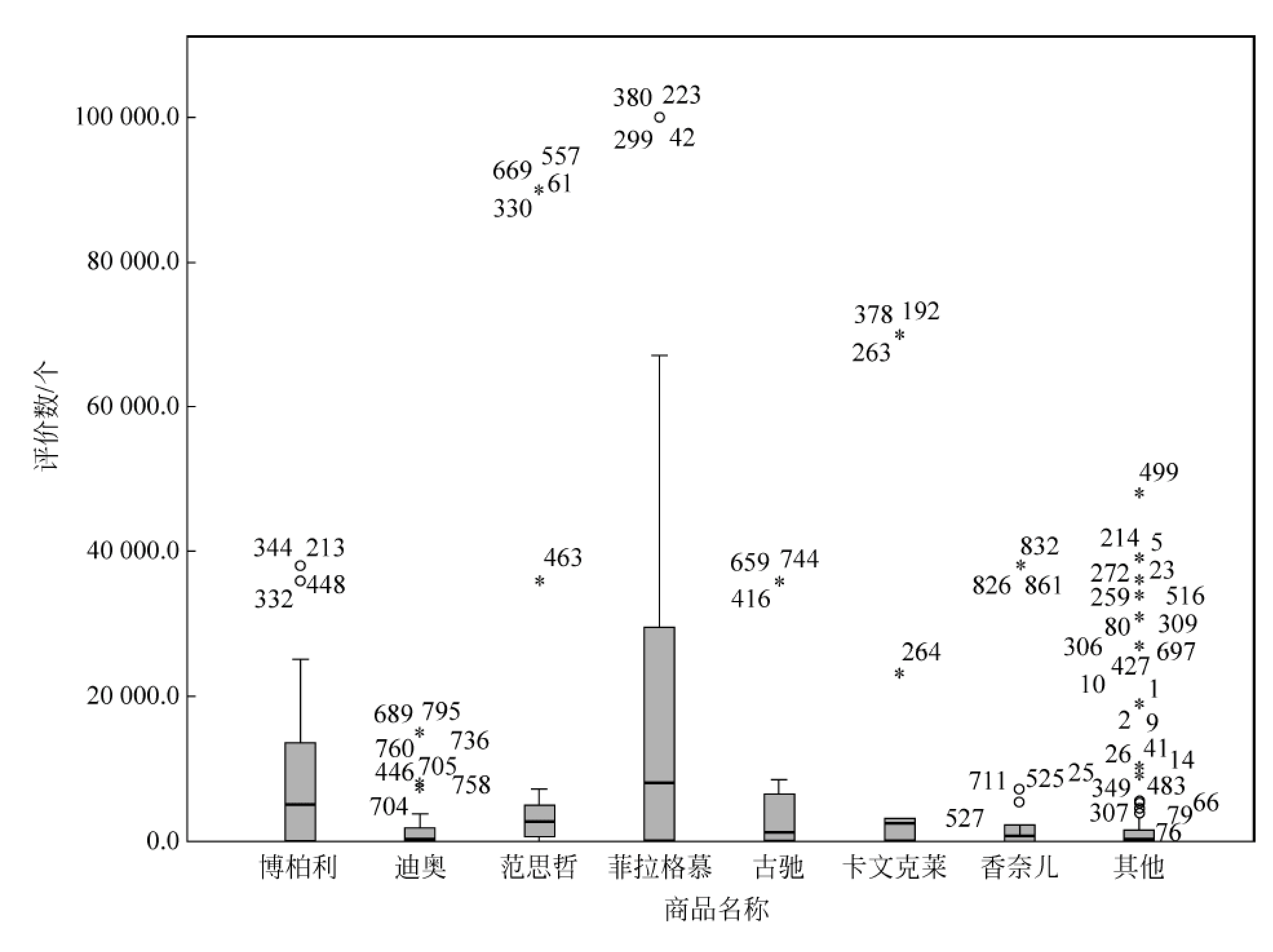


图 3.2 不同品牌香水的评价量箱图

由于箱图本身形状怪异，箱子被压扁且有很多的异常值，因而很难从图 3.2 中得到具体结论。分析其原因，是因为没有对样本数据进行预处理（当然，也不是所有样本数据都需要进行预处理）。针对此类问题，如果数据取值为正数，一个解决方法就是尝试使用对数变换来对数据进行预处理，使幂函数或指数函数的曲线拟合线性化，能够很好地处理不对称分布、非正态分布和异方差等情况。

针对本实例，首先使用对数变换来对样本数据进行预处理。使用底为 10 的对数进行处理，得到评价的对数变换结果，存储至评价数量这一变量中，然后绘制对数变换后不同品牌

香水的评价量箱图,如图 3.3 所示。

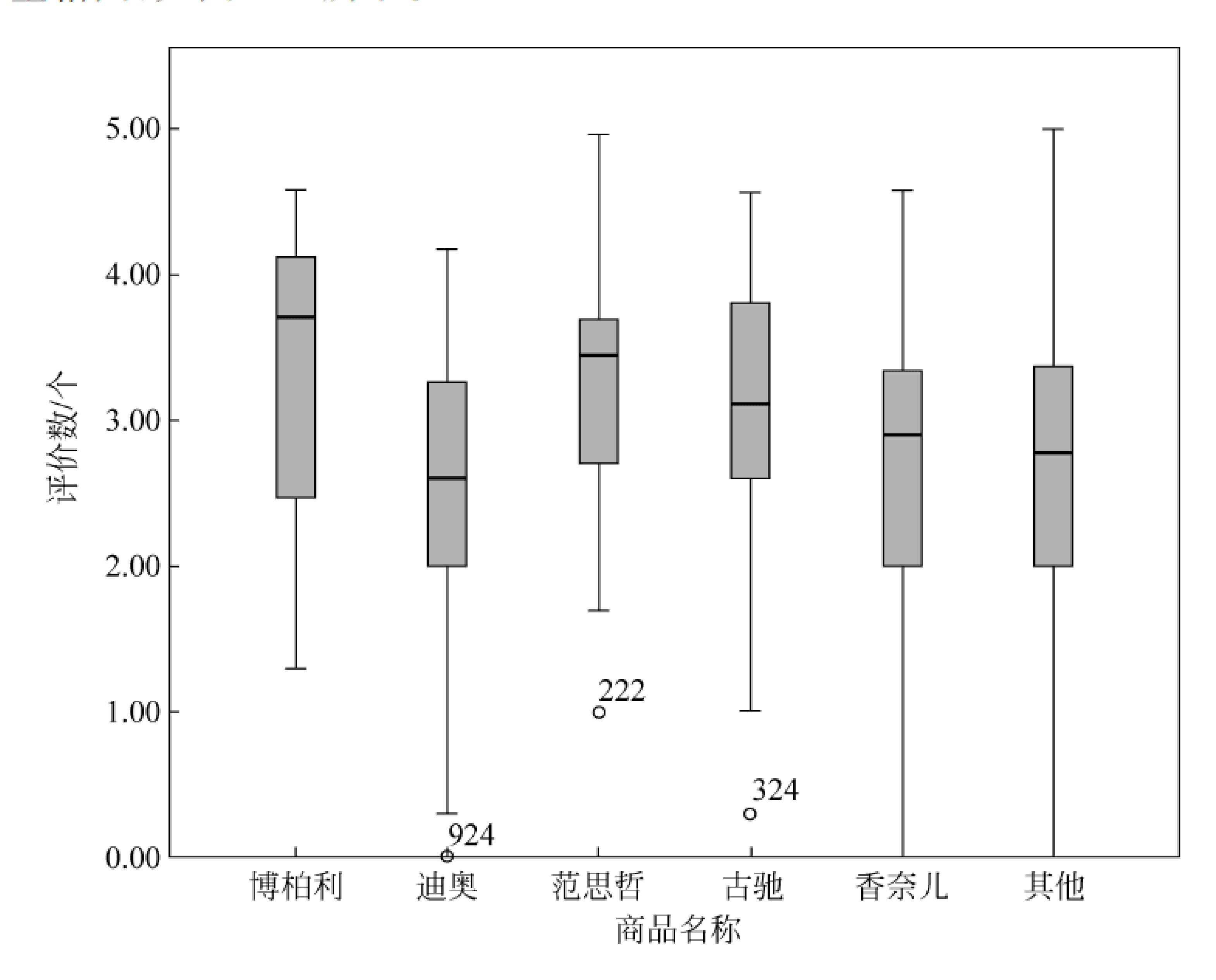


图 3.3 对数变换后不同品牌香水的评价量箱图

对比图 3.2 与图 3.3,发现经过对数变换的箱图可更加直观地表现出数据的平均水平、波动程度和偏态等信息。这一点说明不是所有数据都适合画箱图,但是可以利用数据变换进行预处理,使得数据适合用来绘制箱图。

箱图的另一功能是使用定性变量画分组箱图,各个箱图之间作比较。结合前面分析的实例,图 3.1 只设置了一个定量变量,所以只有一个箱图,而这就让箱图失去了它的一个很重要的功能:多批次数据的对比。而且,一个箱子的箱图是没有必要的,完全可以由直方图来代替。图 3.3 所示箱图设置了定性变量——商品名称,通过商品名称这一定性变量,就能在一个箱图中绘制多个箱子,在同一水平上对各个箱子相关数据节点进行比较,得到多批次数据之间的关系。

3.2 雷达图

雷达图是一种应用于多维数据分析的图形,通过对多维数据进行分析,来探究问题的相关状态。雷达图主要应用于财务分析,其主要作用是将各项数据分析的数或比率,集中展现在一个圆形的图形或者正多边形上,以凸显各种数据比率情况。在财务分析中,雷达图主要用于分析企业经营状况——收益性、生产性、流动性、安全性和成长性的状况。

雷达图可以从静态和动态两个方面分析客户的财务状况。静态分析是将客户的各种财务比率与其他客户或者整个行业的财务比率作横向比较;动态分析是将客户现在的财务比率与以前的财务比率作纵向比较,就可以发现客户财务及经营情况的发展方向和变化。雷达图将纵向和横向的分析比较方法结合起来,综合计算客户的收益性、流动性、成长性、安全性及生产性等 5 类指标。

雷达图的样式是一种类似于蜘蛛网状的图形，主要通过将企业的各种经营比率值连接而成的不规则闭环折线图与各同心圆进行比较，得到企业经营态势的好坏。其中，同心圆最小圆代表同行业平均水平的一半或者最差水平，中心圆中等大小的圆代表同行业的平均水平或特定比较对象的水平，大圆代表同行业平均水平的1.5倍或最佳状态。然后，将同心圆等分为5个扇区，每个扇区指代一个维度，分别代表收益性、安全性、流动性、成长性和生产性指标区域。

上述介绍的财务分析只是雷达图的主要应用领域。在雷达图的通用性方面，其适用范围和规则简要许多，主要解决多维数据的分析，且每个维度都是可以度量的。一般来讲，雷达图的维度数目应大于等于4。

结合香水案例数据，先对数据进行预处理，选取中国、美国、法国、意大利及英国5国，分析其销售量、品牌数、产品质量、平均评价数及平均价格5个维度相关特性，其中，各个维度数据代表其占样本数据相对应维度总体的比例。使用SAP Lumira绘制雷达图，如图3.4所示。

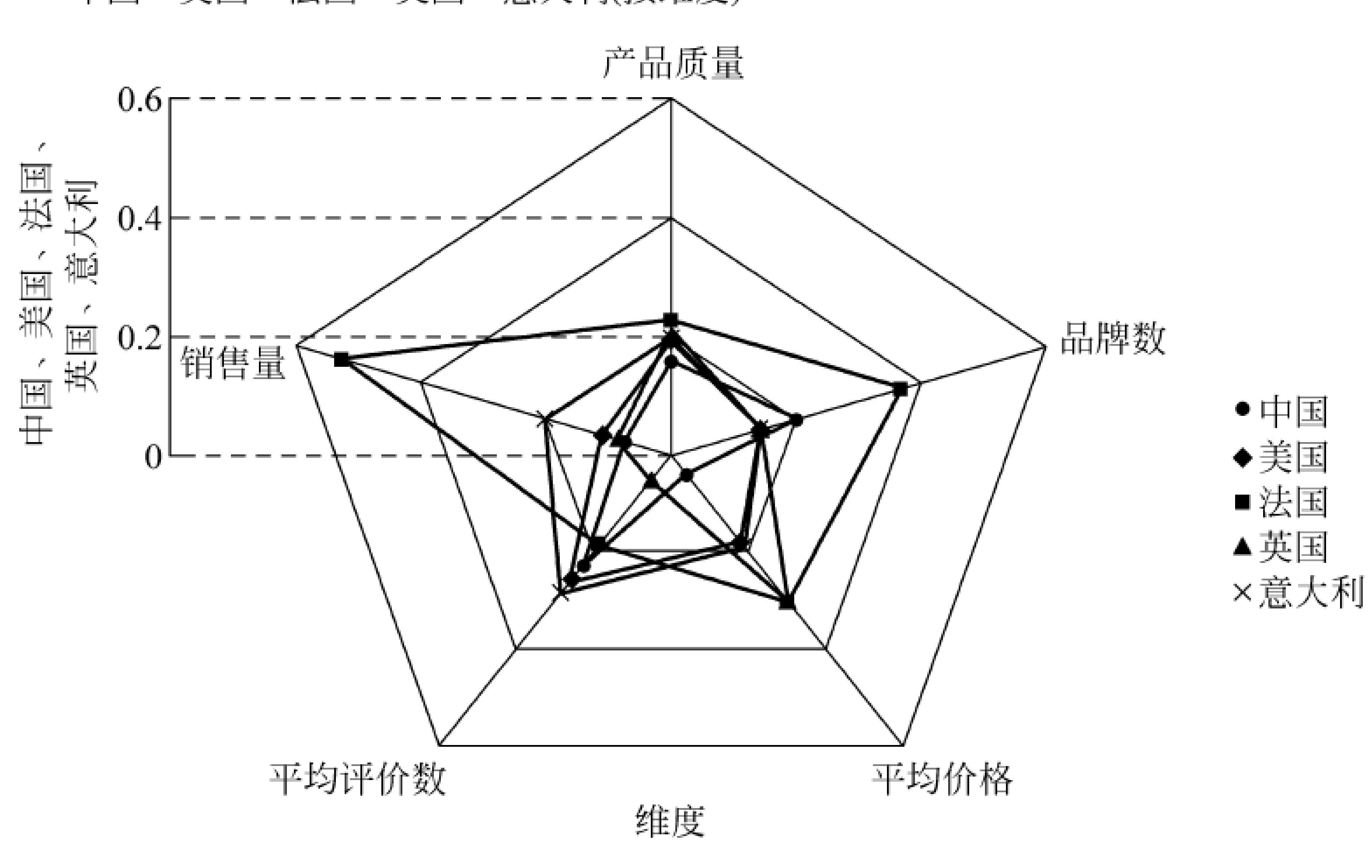


图3.4 雷达图

分析图3.4所示的雷达图，这里的雷达图显然不是应用于企业的财务分析，而是雷达图通用性的体现。此图拥有5个维度，分别为销售量、品牌数、产品质量、平均评价数及平均价格。根据图3.4可以很直观地得到法国在5个维度上基本都占据了优势，因而可以推断出法国在香水领域处于领军地位，而这一点也符合我们对法国香水的认识。美国、英国及意大利3国在各个维度上水平相当，基本都处于中等水平。而中国香水除了在品牌数量之外，其余各个维度基本都处于较低层次水平，因而，中国品牌香水想在香水领域做出一番成就，还有很长的路要走。

综合以上分析，可以总结出以下几条：

- 雷达图主要应用于财务分析，对企业经营状况的5类指标(流动性、生产性、安全性、收益性和成长性)进行评价，来综合评估企业的经营状况。
- 使用雷达图之前，一般需要进行预处理，首先计算出所需分析维度的占比，然后进行

绘图分析。

- 雷达图使用的样本数据，应该由一个或多个定性变量和多个连续型变量组成，因而雷达图能够进行多元多维度比较，更加形象直观地反映出数据特征。如果只是一组连续型变量，则更适合选择箱图来绘制图形。

3.3　标签云

标签云是关键词的视觉化描述，是一套标签和与此对应的权重。标签云作为一种数据可视化工具，其标签一般是独立词汇，通常使用 XHTML(可扩展超文本标记语言，通常设计成一个超链接)元素来表现，按照字典顺序、随机顺序以及热门程度等排序准则进行排序。一般的标签云含有 30～150 个标签，每个标签对应的权重影响标签的字体大小、颜色或者其他视觉效果。标签云是可以交互的，大多数标签都有一个超链接，用户可以单击查看其详细信息。与直方图或饼图相比，标签云能代表更多的信息，尽管不那么准确。一般来讲，标签云有 4 种属性：字号(一般与标签使用次数相关)、排列(按照字典顺序、随机及热门程度等准则排序)、颜色(固定渐进色、是否加背景等)以及字体(可根据自身喜好设置)。

在设计标签云的时候，可以选择比较心仪的图片或者适合用例场景的图片作为填充对象，最后填充得到的标签云就会很形象地展现出来。

运用香水案例数据对数据进行预处理，使用 SAP Lumira 绘制其关于香水品牌、价格和评价量之间关系的标签云，如图 3.5 所示。



图 3.5　香水品牌销量标签云

由标签云的字体属性可知，字体的大小反映了香水品牌价格的高低，即字体越大品牌的价格越高。从图 3.5 能清晰地看出香奈儿、迪奥及古驰等品牌字体最大，因而价格最高。字体的颜色代表了产品品牌的评价量，颜色越深，代表评价数量越多。对此例进行分析，可以

非常清楚地了解各品牌香水评价量之间的比较情况。虽然标签云能形象地表示数据之间的关系，但是其很难得到具体的数据情况。以直方图为例，分析各品牌销量之间的关系。香水品牌销量直方图如图 3.6 所示。

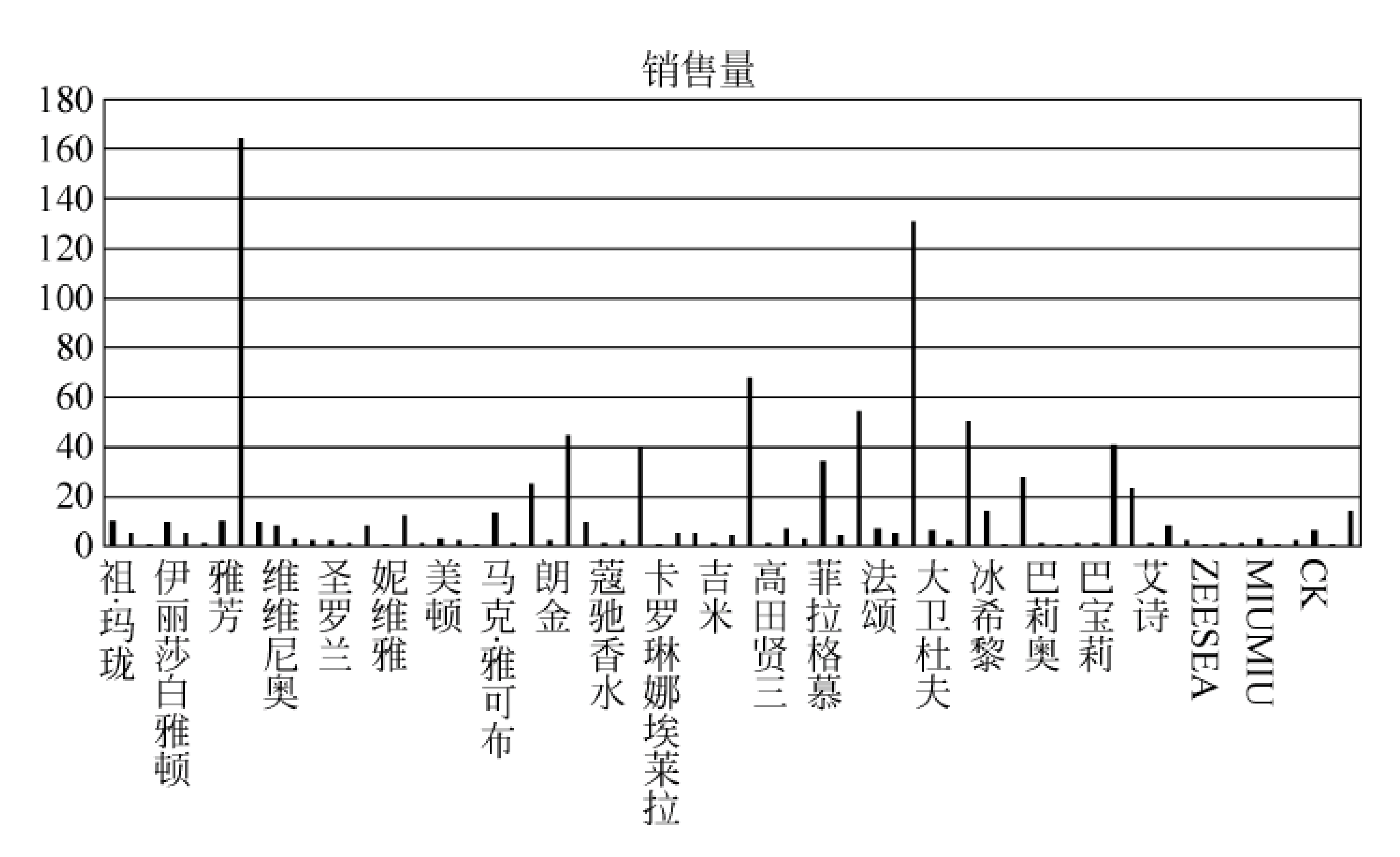


图 3.6　香水品牌销量直方图

对比图 3.5 和图 3.6，在不要求精确性的前提下，标签云比直方图更加形象地展现数据之间的关系，特别是针对权多的数据样本，绘制标签云更有利于对数据进行解读。

综合以上分析，总结如下：

- 标签云使用独立词汇，通过字号、颜色、排序和字体等属性，来形象地体现标签的使用次数及热门程度等相关特性。
- 标签云描述的特性并不能非常精确地呈现出来，因而标签云并不适用于那些对绘图结果要求非常准确的场景。

3.4　气泡图

气泡图中气泡的位置和大小由三维变量决定，其中第一组变量给出直角坐标系的 x 轴值，相邻组变量给出 y 轴值，第三组数据则指代气泡的大小，以上基本就是气泡图的逻辑构成。气泡图基本上与 XY 散点图类似，可以说，气泡图是散点图的升级。散点图只能对成组的两个数值进行比较，而气泡图可以对成组的多个数值进行比较。

使用香水案例数据，并对数据进行预处理，得到不同品牌的平均评价数量和销售数量以及平均销售价格。从中选取香奈儿、迪奥、兰蔻、古驰、范思哲、博柏利、安娜苏、爱马仕和卡尔克莱 9 个占据大多数销售量的品牌进行分析，对数据进行预处理，得到表 3.1。

表 3.1　品牌香水记录表

品牌	平均评价数	平均价格/元	销售量
香奈儿	4303	708	164
迪奥	1882	573	131
兰蔻	3073	388	45
古驰	6262	544	67

续表

品牌	平均评价数	平均价格/元	销售量
范思哲	9359	340	55
博柏利	9511	338	51
安娜苏	5262	289	41
爱马仕	848	717	24
卡尔克莱	12 251	319	40

通过表 3.1,以平均价格为 x 轴,以平均评价数为 y 轴,以销售量来确定气泡的大小,使用 SAP Lumira 绘制气泡图,如图 3.7 所示。

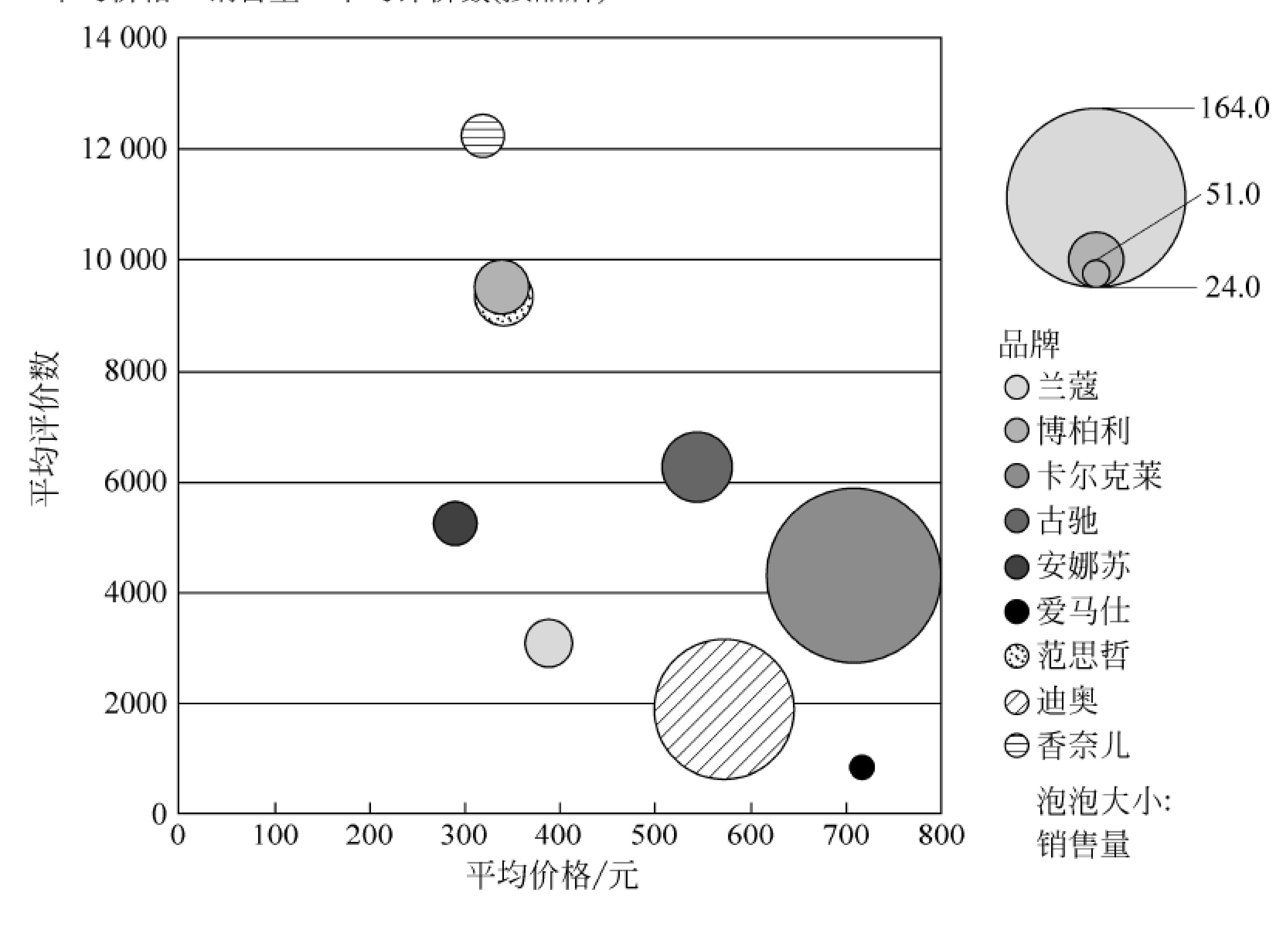


图 3.7　品牌香水气泡图

观察图 3.7 所示的气泡图,通过气泡的位置和大小,很直观地得到各品牌的平均评价数和平均价格及销量在总体样本数据中的具体情况。其中,在本例中,使用不同颜色标记气泡,每种颜色代表一种品牌,这样更利于对比分析各香水品牌的相关数据。气泡的大小反映了销量的多少。可见,香奈儿和迪奥品牌香水销量最高。

3.5　树图

树图是为了达到某种目的或者解决某一问题,采用目的方法或者结果原因方法,层层展开分析,以寻找最好的解决方法或者是查看其根本原因。树图从一个项目出发,展开两个或两个以上分支,然后从每一个分支再继续展开,以此类推,形似一棵树。

按照功能,可以将树图分为两类:对策型树图和原因型树图。对策型树图主要以目的方法方式展开,而原因型树图则以结果原因方式展开。

目的方法方式主要是遵循层层推进规则，对于每一目的，进行层层发问，寻求好的方法或者途径来达到目的，也就构建了对策型树图。结果原因方式则是针对结果进行发问，有哪些原因会导致这个结果或者哪些事项会对这个结果造成影响，通过这种层层推进分析，也就建立了原因型系统图。

树图还可分为矩形树图、组织结构图等。其中，组织结构图用于描述组织结构，一般采用自上而下的展开形式。而对于矩形树图，其主要用来展示层次关系数据。相对于其他层次图表，矩阵树图的优势在于更加直观，并且可以展示层级内的占比关系，直观地反映区域占总体的比率，同时，矩形树图还能依据区域颜色的深浅来反映不同关键词的热门程度。

树图通常用来将主要的类别逐渐分解成越来越详细的层，这样绘制树图有助于思维从一般到具体。在香水案例中，首先对数据进行预处理，得到不同品牌的销售量以及平均价格，并以销售量和平均价格为度量，品牌为维，使用 SAP Lumira 绘制矩形树图，如图 3.8 所示。

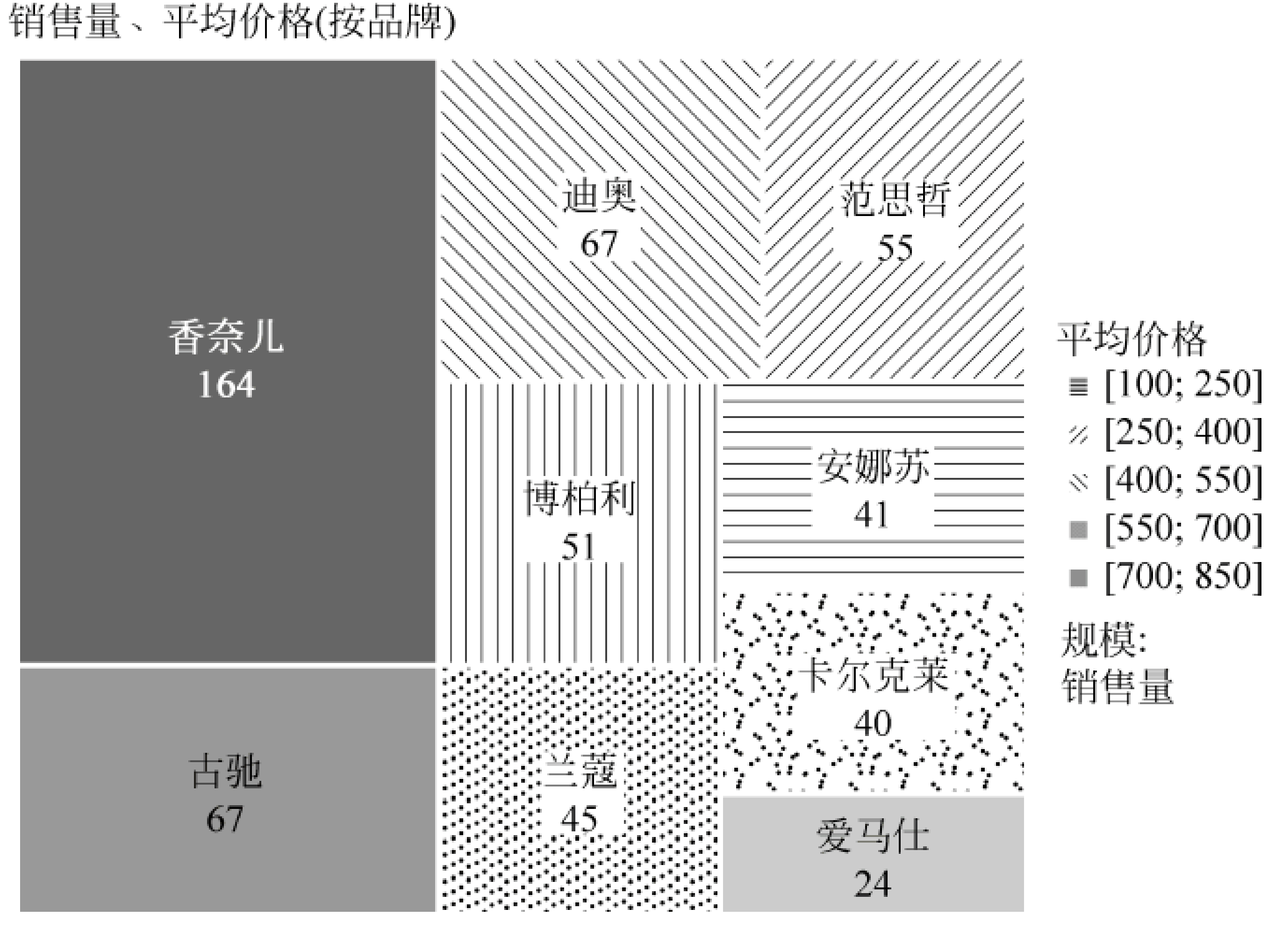


图 3.8 树图

图 3.8 所示树图就是一种典型的矩形树图，其中销售量为第一度量，根据销售量的多少来划分矩形大小。平均价格为第二度量，根据其平均价格的高低决定相应数据模块的颜色深浅。从图 3.8 可以非常直观地看出，香奈儿品牌香水不仅销量好，评价多，而且价格还比较高。这说明了香奈儿品牌香水口碑好，销售情况好。

3.6 地图

地图是以一定的数学、符号化、抽象化法则，使用制图方法，反映客观实际形象的符号模型或者图形数学模型。地图是按照一定的比例运用符号、颜色、文字注记等描述显示地球表面的自然地理、行政区域、社会经济状况的图形。在适用于数据分析的地图图形前提下，这

里介绍的是数据地图。遇到数据与地名的场景时，使用地图能更加形象地展示数据与地理位置之间的关系。数据地图是一种将数据与地理信息有机结合的一种地理数据表达方式。数据地图是以图形化的方式分析和展示与地理位置相关的数据，使得数据与地理之间的关系更加直观、形象化。地图能够很直观地反映数据与地理位置之间的关系，且这种关系可以是分层的，通过地图下钻操作，可以探查数据与不同层次地理位置之间的关系。同时，还能通过颜色的深浅来判断地区或关键词的热门程度。

地图图层是对空间表达的一种重要途径。一个地图可以拥有多个图层，将其叠加就能得到地图的底层（类似于背景图层），构成地图中最基本的地形、地貌数据及某些附属数据或信息。

建立地理智能对象时，如果将地理位置信息转换为地理层次结构，建立地理层次结构，就能实现地图的下钻操作。在香水案例中，以商品产地为例，对数据进行预处理后，建立地理层次结构。以评价量为度量（评价量能在一定程度上反映商品销量），使用 SAP Lumira 绘制包含下钻操作的地图。在地理维度里会出现分层次的地理维度选项，可以选择以国家或地区为维度标准，也可以选择以城市为地理维度标准，得到以城市为维度的地图。

地图下钻能够很好地处理地理位置之间具有层次关系的问题。在进行下钻操作后，可以将研究问题进行细化，分析问题的局部特点。例如，通过对中国香水销售的分析可以了解到中国香水的几个产地，以及该产地香水评价量的平均水平，进而推算出各产地销售量的大概水平。

3.7 高低图

高低图是采用多条垂直线段表示数值区域的统计图形，能够将数值区域形象地表示在图形上，通过多组数据并行比较，易于分析数据区域的相关特性。高低图与折线图、散点图、条形图等统计图相比，它既有研究数据长期变化的特性，也有研究短期内数据变化的特性。因为这些特性，高低图广泛应用于股票、商品、货币及其他市场数据分析中。

高低图绘制过程中，纵坐标表示一个三维数据，分别为高值、低值及关闭。高值代表对应数值区域的上限，低值代表对应数据区域的下限，关闭是用户指定的一个特殊变量（如数据集的平均值），可以在数据区域上以小圆圈的形式标出。

为了更加直观地了解高低图，依据香水样本数据，以香水价格平均数为关闭变量、香水价格最小值为低值变量、香水价格最大值为高值变量，以香水品牌确定类别，使用 SPSS Statistics 工具绘制高低图，如图 3.9 所示。

图 3.9 显示了一些主要的香水品牌价格的高低图，从中可以直观地了解各数据集的区域范围。通过关闭在高低图中小圆圈的表示，可以了解各品牌香水的平均价格。通过对高值变量的分析，可以得到迪奥和香奈儿两个品牌香水基本占据了高端香水市场。各品牌香水的平均价格在 400～700 元波动。

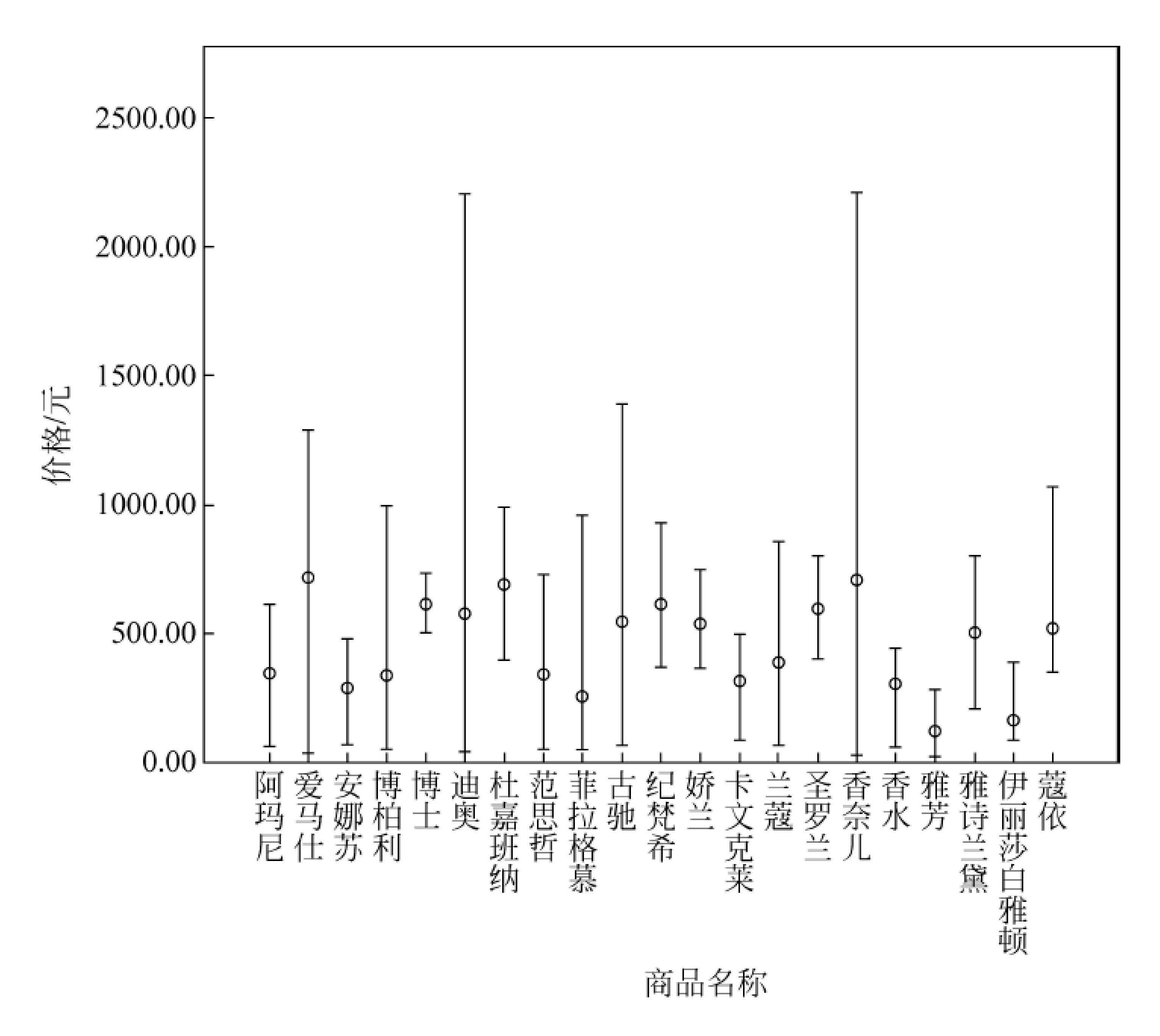


图 3.9 高低图

3.8 双轴图

双轴图是一种单 x 轴双 y 轴的统计图形，这里的双 y 轴指有两条 y 轴，位于两侧，代表了两种变量含义，这样就能根据 x 轴变量变化绘制出两条不同图形，可以简易地看成两种统计图形的合并。绘制的图形可以采用多类图形，如折线图、直方图和散点图等。

双轴图可以依据 x 轴的分类类型分为两类：包含分类 x 轴的双 y 轴及包含刻度 x 轴的双 y 轴。分类 x 轴的双 y 轴类型对应的 x 轴变量是非连续型变量，而包含刻度的 x 轴的双 y 轴类型对应的 x 轴变量是连续型变量。

双轴图能够在同一统计图形上采用两种绘图方式，并且将结果展示在同一统计图形上，使得能够更加形象地对比分析多组数据特征。当采用不同统计图形的绘制方法绘制双轴图时效果对比明显，如折线与直方图的叠加，就能够很好地展示数据特征。

为了更加形象地表示双轴图的特征，依据香水案例，设置香水品牌为分类 x 轴，左侧 y 轴为各个品牌评价平均数，右侧 y 轴为各个品牌价格平均数，使用 SPSS Statistics 工具绘制的双轴图如图 3.10 所示。

通过图 3.10 所示的双轴图，在已列的各品牌香水中，香水的平均评价数高的对应的平均价格一般较低，香水的平均评价数低的对应的平均价格一般较高。由此可见，不同品牌香水的平均评价数与平均价格呈现一种负相关趋势。对于那些平均评价数中等的品牌，其平均价格也保持中等。

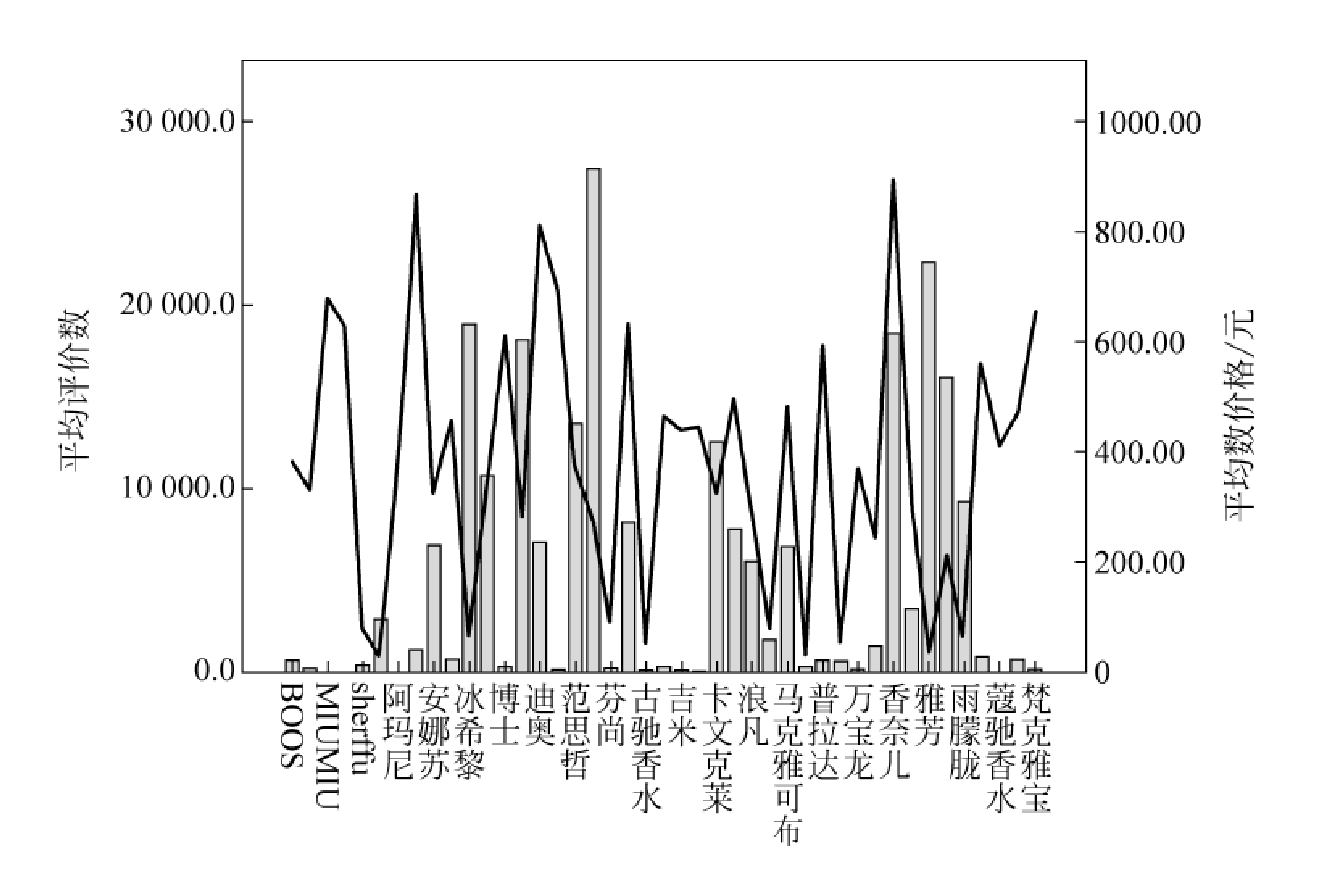


图 3.10　双轴图

3.9　关系图

关系图是使用连线的方式，将相关事务连接起来，表示事物相关性的图形，这里的相关性指的是事物之间复杂的逻辑关系。关系图通过对事物相关性的研究，找出事物之间复杂的逻辑关系，从而找出要素和解决问题的方法。

在对事物的逻辑结构进行分析时，如果分析的关系是纵向关系，即分析的是单个事物内部因素之间的关系，则可以选用“原因结果、目的方法等”方法来研究。如果分析的关系是横向关系，即需要分析多个事物之间的复杂逻辑结构，这时就需要应用关系图。

关系图按其应用目的来分，可分为：单一目的型和多目的型(研究问题的个数)；按照其分布结构来分，可分为：中央集中型(箭头向内集中)和单向汇集型(箭头单向顺延)。在对关系图分析时，箭头只进不出的是问题；箭头只出不进的是重要因素；箭头有出有进的是中间因素。需要指出的是，一般适用于关系图的场景应该是事物之间的逻辑关系非常复杂的，简单的场景可以由树图等来研究其相关逻辑特征。

为了更加直观地了解关系图的相关特征，结合香水案例，对商品名称、香调、分类、商品产地及性别 5 个字段使用关系图来了解其相互逻辑关系。使用 IBM SPSS Modeler 绘制关系图，如图 3.11 所示。

分析图 3.11 所示的关系图，关系图中各元素字段之间使用线段粗细来表示两因素之间联系的强弱，线段越粗、颜色越深，代表其联系越强。这里给出了其链接强度的相关数据，见表 3.2。

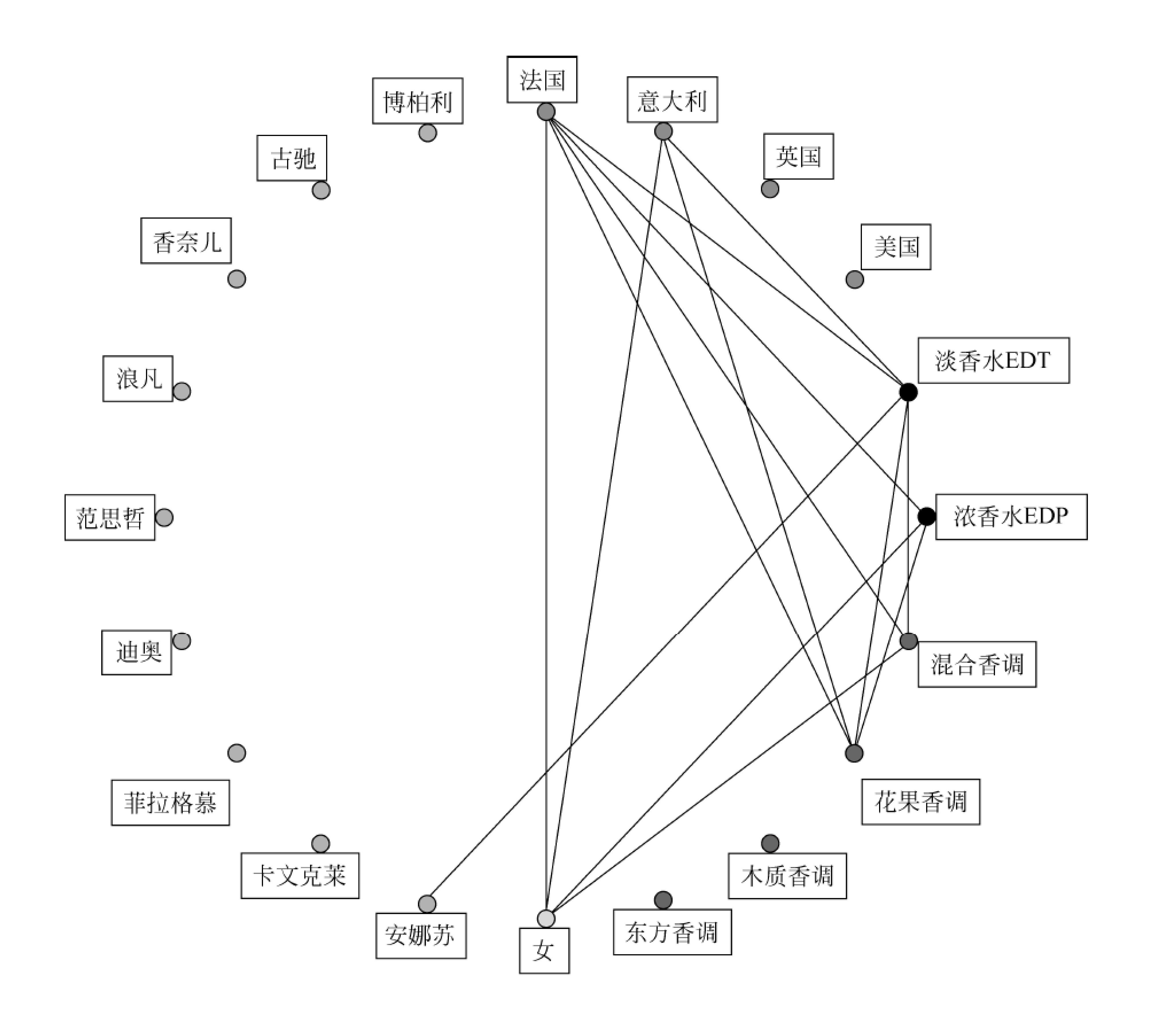


图 3.11 关系图

表 3.2 链接强度表

链接	字段 1	字段 2
6.65%	香调="花果香调"	性别="女"
5.98%	分类="淡香水 EDT"	性别="女"
4.66%	商品产地="法国"	性别="女"
4.62%	香调="花果香调"	分类="淡香水 EDT"
3.27%	分类="浓香水 EDP"	性别="女"
3.21%	香调="花果香调"	商品产地="法国"
2.76%	分类="淡香水 EDT"	商品产地="法国"
2.12%	香调="花果香调"	分类="浓香水 EDP"
1.99%	分类="浓香水 EDP"	商品产地="法国"
1.92%	商品产地="意大利"	性别="女"
1.67%	香调="混合香调"	性别="女"
1.58%	分类="淡香水 EDT"	商品产地="意大利"
1.54%	香调="花果香调"	商品产地="意大利"

续表

链接	字段 1	字段 2
1.03%	商品名称="古驰"	性别="女"
0.98%	香调="混合香调"	分类="淡香水 EDT"
0.98%	香调="混合香调"	商品产地="法国"
0.94%	商品名称="博柏利"	商品产地="法国"
0.92%	商品名称="博柏利"	性别="女"
0.85%	商品产地="美国"	性别="女"
0.83%	商品名称="古驰"	商品产地="法国"
0.79%	商品名称="范思哲"	性别="女"
0.77%	商品名称="博柏利"	香调="花果香调"
0.75%	香调="花果香调"	商品产地="美国"
0.75%	商品名称="范思哲"	商品产地="意大利"
0.75%	商品名称="古驰"	香调="花果香调"

对照表 3.2 和图 3.11，可以看出香水香调为花果香调、香水分类为淡香水 EDT、性别为女的链接关系占据高位，法国的香水也与女性的链接关系比较高。这表明了女性用户更加青睐法国产花果香调淡香水 EDT 类型的香水，而且花果香调的香水与淡香水 EDT 类型更加搭配。

3.10 热图

热图是一种表现数据热点的图形，以区域和颜色等视觉效果，形象地表现数据的密度、频率及热点等特征。热图的热字体现在图形表达时，其数据热度等信息一般以火焰色彩表示，展现出极强的视觉表达力。

热图是以区域颜色深浅效果来展现数据特征，因而，热图表达的仅仅是数据之间的大概关系，并不能精确展现数据频率、热度等特征。

热图可以看作是地图的增强版。地图展现的是数据与地理位置之间的关系，而热图则是将地理位置广义化，以不同的区域块来区别。

在 Web 领域，热图被用来检测页面的哪些部分对顾客具有吸引力。其原理是：记录用户点击的区域，以热图的形式展现出用户得到点击区域，然后根据具体情况对页面设计进行调整，设计出更具用户友好型的网站。在其他领域，热图也有很广泛的应用，天气预报中的气温图就是典型的热图。

下面结合香水案例来了解热图。香水的评价量在一定程度上反映香水的销售量，以香水平均评价为度量，以商品名称及香调为维度，使用 SAP Lumira 来绘制热图，如图 3.12 所示。

分析图 3.12，该热图表现的是香水品牌的各香调类型香水的评价量（销售量）分布热点。各香水品牌的香调类型主要是花果香调和混合香调类型，其平均评价量热点分布主要是 0～15 000 的水平。其中，卡文克莱及香奈儿拥有大多数香水香调类型，且香奈儿香调类型香水的平均评价较高，而卡文克莱的平均评价较低，反映了香奈儿香水的销量高，卡文克莱的销售量较低。

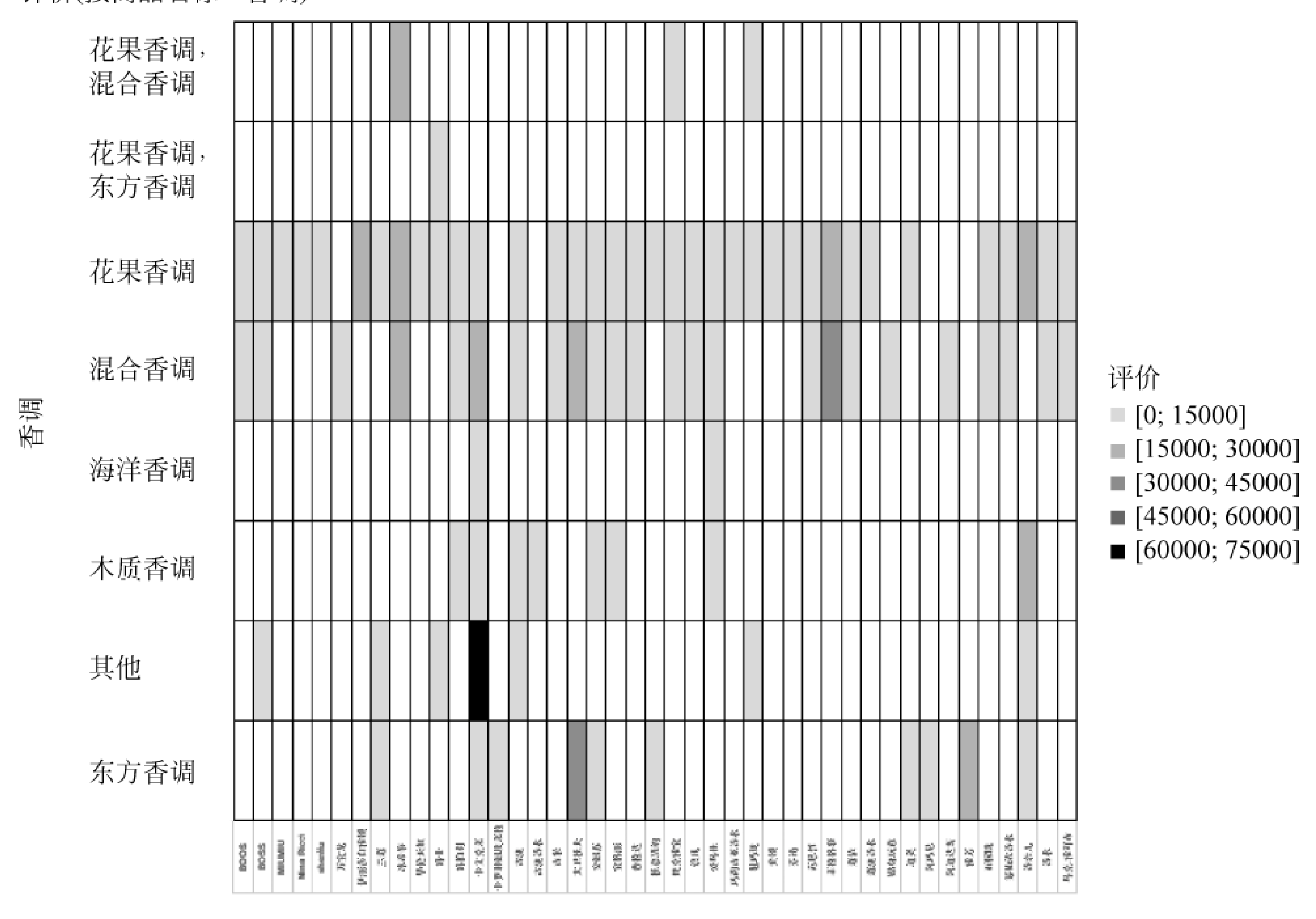

评价(按商品名称、香调)
香调
花果香调，混合香调
花果香调，东方香调
花果香调
混合香调
海洋香调
木质香调
其他
东方香调
商品名称
评价
[0; 15000]
[15000; 30000]
[30000; 45000]
[45000; 60000]
[60000; 75000]

图 3.12 热图

第 4 章

SPSS Modeler建模组件介绍

数据挖掘可以理解为从数据中挖掘出正确的、有用的、未知的、综合的知识。IBM SPSS Modeler 包含数据获取、数据预处理、数据建模、评估和部署等一系列步骤的组件,分析人员可通过拖放的方式组合节点完成数据挖掘流程。

由于业务问题的提出与理解需要数据分析师人工分析,目前数据挖掘软件还很难自动化,因此本章主要介绍数据挖掘工具 IBM SPSS Modeler 常用组件。这种工具基本不需要编程(18.1 版本补充了 Python 组件,与 18 版本的 R 组件一起可以方便用户编程实现一些分析功能,提高了挖掘系统的灵活性),比较适合不喜欢编程或不会编程的数据爱好者,尤其是业务人员。但要熟练使用 IBM SPSS Modeler,也需要了解常用的统计操作和数据挖掘算法,这样才可能选择合适的组件搭建数据挖掘流程。

4.1 数据预处理组件

数据预处理对数据挖掘的效果起着至关重要的作用。好的数据预处理能为数据挖掘提供高质量的数据,是数据挖掘成功的重要保证,但现实的数据中往往存在不完整、异常和不一致的记录,这就对提高数据质量提出了很高的要求。数据质量包括准确性、完整性、一致性、时效性、可信性和可解释性,在对数据预处理的过程中,需要紧紧围绕上述要求展开。

在实际的数据处理中,我们对数据的清理、集成、选择、变换并没有严格区分,更多是为了逻辑和思维上的清晰来对节点进行分类。在实际业务处理中,往往是各种处理技术混合使用,并没有严格区分。

4.1.1 数据清理组件

数据清理包括填补空值,剔除噪声,识别、删除离群点。其重要性在于,如果数据是“脏”

的，则在学习的过程中，会使模型向错误方向倾斜，丢失重要信息，甚至完全陷入混乱，并且可能挖出完全没有意义的知识，甚至最后出现无效的、错误的结论。

1. 区分节点

区分节点的作用是去除数据库中重复的数据。废弃重复记录的第一个记录，将部分重复的数据扔到数据流中。

区分节点的使用方式有3种：

(1) 为每个组创建一个组合记录。可以根据需求指定如何分组，如何创建组合记录。

(2) 仅包含表头。在一些数据中，常常有一个汇总的表头，删除其余的详细数据。

(3) 删除表头。对数据进行细微分析时，删除汇总的表达，对数据的内部结构进行仔细分析。

(4) 关键字段区分，设置一些关键字，当数据中有这些关键字时，对此数据进行区分。

2. 填充节点

填充节点是按条件补充数据和存储类型。可以用一些特定的规则来替换特殊值或者空值。例如，用0值填充NULL。

填充节点通常和类型节点组合使用，以替换空值。其用法：使用字段选择器从数据中选出要查找的字段，然后在替换选项中选择替换方法。可以是基于条件的，根据指定条件，进行替换。

3. 过滤节点

过滤节点可过滤多余字段数据，并在此节点对数据属性等进行一些更改，使数据更“干净”，提升数据质量和建模效率。

过滤选项中使用的表格，可以呈现每个字段进入和输出节点时的名称，可以对重复的或者不需要的字段进行重命名或者删除。例如，对用户编号、产品编号等无意义字段进行过滤。

4.1.2 数据集成组件

数据集成指合并来自多个数据存储源的数据，有助于减少数据的重复和不一致，从而提高数据的质量，并优化模型的准确性和运算效率。其中，数据集成还涉及数据值冲突时的检测与处理。

1. 汇总节点

汇总节点是对记录(行)进行操作的节点，作用是对各字段进行加总、合计、取均值等操作。处理汇总操作后可以增加新的字段，但是在汇总之前要对记录进行预处理，将缺失值进行处理，否则对汇总结果造成影响，最终分析结论产生较大误差。

2. 合并节点

合并节点的作用是合并多个输入数据，并输出包含某些关键字段数据的输出。合并节点广泛使用在不同数据源的合并和集成，避免重复数据。

合并节点一般有3种方式：①按关键字合并，按关键字编号进行内连、外连等合并；②按指定条件合并，在节点设置中设置相应的条件，对满足条件的数据执行合并；③按重要

性合并,例如,在数值累加的情况下,一些对结果影响过小的数据,在合并中适当忽略。

3. 追加节点

追加节点是将一个源中的数据传递到下游流程中,作用是连接各组记录,合并类似结构的数据,所以,各源的字段类型需要一致,即分类类别无法追加到连续字段中。如果是数据结构不同的数据集,则没太大作用。

4.1.3 数据选择组件

数据选择可以用来得到数据集的简化表示,虽然数据容量上小得多,但是能够保持数据的完整性,规避数据冗余,并产生同样的分析效果。

1. 选择节点

选择节点可以从一些数据库(或数据流)中,根据特定的某个条件,选择出一些符合特定要求的数据,独立地呈现在输出中。

选择条件可以用 CLEM 进行指定,在窗口中可以输入函数,来选择符合条件的数据。例如,选择 value1 的条件:(var='value1')。再如,删除空值的数据的条件函数:not (@NULL(var1) and @NULL(var2))。

2. 样本节点

样本节点可对庞大的数据进行抽样,用于提高计算性能和选择对应数据进行专门分析,以此提高效率。

其优势是在条件允许的情况下,对抽样样本评估可以提高运行效率;可以选择特定的记录或者交易组进行分析。例如,对偏离值分析,或者对购物车的分析;可以对指定数据或者观测值进行随机数据分析。

4.1.4 数据变换组件

在数据预处理中,数据通常被变换或者统一格式,使挖掘过程耗费时间更短,更有效,更精确。通常,数据变换的方式有以下几种:光滑、构造特征、聚集、规范化、分门别类。

1. 类型节点

类型节点是非常重要的节点,其作用是对指定的字段元数据和相应的属性进行更改,可以对数据的测量级别和属性进行过滤、修改。此外,还可以设置控制选项、字段建模、制定标签、指定值等。

2. 平衡节点

平衡节点主要针对特别分散的数据,可以遵循指定的系数条件,调整数据不集中的比例。平衡是通过复制记录或随机删除的方法来实现的,所以,每次运行其结果集并不固定。一般要选中“仅平衡训练数据”,特别是在遇到不平衡检验或验证分区得分时,当然,如果流中不存在分区字段,则此选项无效。

3. 导出节点

导出节点的作用是修改、创建新字段,导出的形式包括标志、状态、条件、计数和公式等,

可以导出单个或多个字段。

导出多个字段时，可以应用@FIELD 函数，并结合公式在不指定固定字段的情况下同时生成多个字段，并且可指定新字段的命名规则。

4. 分级节点

分级节点的作用是将连续数值离散化，将连续变量字段的值自动分级成多个新的分类字段。此种方法在决策树中应用广泛，例如，成绩是连续变量，应用分级节点，把成绩分为优秀、良好、中等、较差、很差等多级。

一般在以下几种环境中应用分级：

(1) 模型性能。针对一些对连续变量处理效果不佳的算法模型，非常有必要运用分级节点，如决策树、Logistic 等。运用分级节点，会大大增加模型准确度。

(2) 算法要求。某些特定的算法要求将输入进行分类，如朴素贝叶斯、逻辑回归。

(3) 数据隐私。敏感信息，保护个人隐私，如收入、身份证号等。

4.2 数据挖掘建模组件

数据挖掘模型是一系列规则、公式或方程组，可以根据多组输入或者变量来预测输出或者进行分类、聚类、关联或者回归等分析。在何种情况下选择何种数据挖掘模型至关重要。SPSS Modeler 中集成了目前主流的数据挖掘方法，可以很方便地依照数据特点生成预测模型，并将其应用于商业活动中，从而改进商业决策过程。

4.2.1 模型筛选

在数据预处理完毕之后，通常还要做一步数据准备工作，即特征选择和异常检测，按照重要性对变量进行排序，并将异常数据剔除。特征选择和异常检测可以组合使用。

1. 特征选择

在数据挖掘时可能会遇到大量的备选变量，要花费大量的时间和精力来对这些变量进行分析，使用特征选择算法来识别重要的变量就变得尤为重要。通过把注意力集中到最重要的变量上，可以忽略无效计算，有效降低计算量，加快计算速度，并提高运算效率。

2. 异常检测

异常检测节点的作用在于发现离群点，针对那些不符合正常规律数据模式的离群点进行确认。异常检测通常是检测大量变量后，识别相似记录所属的聚类或对等组，然后将数据与组内数据进行比较，以识别异常值。观测值离聚类中心越远，越有可能是异常点。

4.2.2 自动建模

自动建模的节点可以对多个算法进行自动评估和比较，按照某种结果评价标准进行排序，减少用户的手动操作工作量，可以快速验证多种模型，并可对某一模型进行参数设置和选项配置。

1. 自动分类器

自动分类器节点使用多种不同的方法来估算和比较目标模型，可以在一次建模运行中尝试多种方法。可以选择所用算法，并试验选项的多个组合。例如，无须为神经网络选择快速、动态或修剪中的某个方式，完全可以全部尝试。该节点将探究每种可能的选项组合，并根据指定的测量对每个候选模型进行排序，然后保存最佳模型，以用于评分或进一步分析。

2. 自动数值

自动数值节点使用多种不同方法来估算和比较模型，以得出连续数值范围结果，可在一次建模运行中尝试多种方法。可以选择所用算法，并试验选项的多个组合。例如，可以使用神经网络、线性回归、CART 和 CHAID 模型预测住房价值，以确定哪种模型的性能最好，并且可以尝试步进、向前和向后回归法的不同组合。节点研究选项的每个可能组合，根据指定的测量为每个候选模型排序，并保存最佳模型，用于评分或将来的分析。

3. 自动聚类

自动聚类节点是通过评估和比较来识别具有类似特征记录组的聚类模型。节点的工作方式与其他自动建模节点相同，可以在一次建模运行中试验多个选项组合。模型可使用基本默认参数进行比较，以尝试过滤聚类模型的有效性以及对其进行排序，并提供一个排序依据，如轮廓、聚类数、最小(大)聚类大小、聚类大小、评估字段重要性等。

4.2.3 决策树模型

决策树模型是指可以根据可解释的决策规则，对未来的观测值进行预测或分类的分类系统。其模型的优点有：

(1) 决策树推理的模型具有非常清晰的逻辑。

(2) 决策树推理的模型只具有真正影响决策的属性。

(3) 区分方式可以转换成 IF-THEN 规则集合。

(4) 可以观察出如何根据属性将总体分割或分区成相关子集。

1. 分类和回归树

分类和回归树节点(CART)，可以用于预测或分类未来观测值的决策树。CART 算法使用了一种 Gini 指数(不纯度函数)来度量数据集中度，首先计算各个属性的纯度增量，然后选取纯度增量最大的属性，拆分数据集，所有分割均为二元分割。

2. CHAID 节点

CHAID 节点(Chi-squared Automatic Interaction Detector)使用卡方作为度量统计学显著性的方法，较高的卡方值表示用某属性拆分决策树时，可以把样本集拆分为有显著分布差异的分组。

算法的核心思想是：根据结果变量与解释变量对样本进行最优分割，按照卡方检验的结果进行多元列联表的自动判断分组。

CHAID 算法的优势是可以生成非二元树，即有些可以分割成多于两个的分支。因此，与二元生成方法相比，CHAID 倾向于范围更广的树。CHAID 适用于所有类型的输入变量，并能接受权重和频率变量。

3. QUEST

QUEST是Quick Unbiased Efficient Statistical Tree的缩写,是用于构建二元分类的决策树。设计目的是减少大型CART分析所需的处理时间,也减少发现的趋势,以便支持多个分割的输入,所有分割都必须是二元的。

4. C5.0决策树

C5.0的核心思想是:根据每个级别提供最大信息增益的字段分割样本,在分割节点的选择上可以进行多次多于两个的分割。C5.0算法的优势是:缺少数据以及存在大量输入字段等问题时,C5.0模型的表现十分稳健,且不需要花费过长的时间。模型的结果更易于理解。此外,C5.0还有增强方法来提高分类的准确性。

5. 随机森林

随机森林节点是一种基于树的分类和预测方法,此方法根据分类和回归方法构建。与CART类似,此预测方法使用递归分区将训练记录拆分为具有相似输出字段值的段。首先,此节点通过检查可供其使用的输入字段来查找最佳分割。分割可定义两个子组,其中每个子组随后又分割为两个子组,以此类推,直到触发其中一项停止标准为止。所有分割都是二元的(仅有两个子组)。

4.2.4 贝叶斯网络模型

贝叶斯定理是一种把先验知识与样本中得到的新信息相结合的统计方法。贝叶斯网络是一种基于贝叶斯定理的图形模型,可以显示数据集中的变量以及概率,还能显示这些变量之间的条件和独立性。

选用贝叶斯网络可以了解因果关系,可避免过拟合,可清晰观测到逻辑视图。常用的结构有TAN和马尔科夫覆盖。

TAN:树结构朴素贝叶斯模型是简单的贝叶斯网络模型,该模型除了随目标变量变化外,还随其他预测变量变化,因此增加了多维因素的准确率。

马尔科夫覆盖:常用于数据集中的节点的集合。马尔科夫覆盖基本上包含了与预测目标变量相关的所有变量,上一维和下一维。但是,当处理大规模数据集时,会由于变量过多,耗费过多的处理时间。

4.2.5 神经网络模型

类神经网络是模拟人类的神经元结构,形成输入层、隐藏层、输出层,来模拟人脑处理信息的简易模型,模拟大量类似于神经元的物质互联处理信息。自变量和因变量之间的关系是在模型学习的过程中确立的,可以是线性的,也可以是非线性的,非常灵活,其缺点是对规则的解释性较差。

神经网络包含3层:输入层,输入外界信息,影响因素的变量;隐藏层,经过输入层的权重变化之后的函数变换,再变化权重至输出层,类似于大脑的思考过程,判断哪种因素更重要;输出层,即需要预测和判断的目标。网络通过不断地学习过程,当预测结果与样本

目标不一致时，调整权重。

4.2.6 支持向量机模型

支持向量机(SVM)是一种分类和回归技术，应用非常广泛，特别是，在小样本集中也能得到较好的结果，其特点是自变量数目较多时也不会出现维数灾难。

1. 支持向量机

支持向量机具有坚实的统计学理论基础。它的核心思想是用两类线性可分问题来说明可以找到一个超平面，该超平面可以把训练样本分为两类。分类间隔是离超平面最近的样本，且平行于最优超平面的两个超平面。

2. 线性支持向量机

线性支持向量机(LSVM)节点可以使用线性支持向量机对数据进行分类。LSVM 特别适用于大型数据集，即具有大量预测变量字段的数据集。可以对节点使用默认设置，以便相对较快地生成基本模型，也可以使用构建选项来试用不同的设置。LSVM 节点类似于 SVM 节点，但它是线性的，更擅长处理大量记录。

4.2.7 时间序列模型

时间序列是指在不同时间上的观察值或事件组成的序列。时间序列建模方法假定历史会重演。即使不完全一样，也会非常接近。

1. 时间序列

时间序列的数据通常具有以下几种特征：趋势、周期运动、季节性变化、不规则运动。“时间序列”节点可以在本地或分布式环境中与数据配合使用；在分布式环境中，可以利用 IBM SPSS Analytic Server 的能力。通过此节点，可以选择对时间序列的指数平滑法模型、单变量自回归积分移动平均值(ARIMA)及多变量 ARIMA(或转换函数)模型进行估计和构建，并根据时间序列数据产生预测。

2. STP 模型

STP 是时间-空间预测的缩写，通过对时间和空间中的指标进行测量，来分析预测某一时间点的指标值，模型需要指定位置数据、输入变量、时间变量和目标字段等数据，其中目标变量只能为连续型变量，位置类型只能为地理空间字段，时间变量要预处理为具有固定间隔的索引字段，也可以在模型的时间区间中指定。

4.2.8 统计模型

统计模型使用统计分析方法，从数据中挖掘有用的信息。在样本够大和特定的情况下，统计方法可以非常快速地给出合适的模型。

1. 线性模型

线性模型通过构建目标变量和预测变量之间的线性关系来预测连续目标的变化。线性模型相对简单，用于评分的数学公式也易于解释。与其他模型类型相比，其属性易于理解，

并且可以进行快速构建。

2. 回归模型

回归模型是一种对统计关系进行定量分析的模型，用于构建目标变量和预测变量之间的关系，线性回归通过拟合预测输出值与实际输出值之间的差异最小化直线或者曲面，来达到预测的目标。此模型的优势是形成预测的数学公式易于理解，且训练速度非常快。

模型中的变量只能为数值型变量，如果目标字段为非连续型变量，可以使用逻辑回归来代替。模型构建方法中的进入法不对输入字段作任何处理，步进法、后退法、前进法将对输入字段进行优化。

3. 逻辑回归模型

逻辑回归是一种常用的统计学方法，是根据输入值对记录进行分类的统计方法，相比于线性回归，其分类目标是可以为类型字段，按目标变量类型可应用二项式或多项式算法。

逻辑回归的优点是：通常模型比较准确，可以处理符号或者数值型数据。可以完整地给出所有目标的预测概率，以比较次优选项。处理超大型数据集时，可以禁用高级输出选项，如似然比检验等，改用 Wald 统计量和评分统计量，从而提高性能，减少建模时长。

4. 主成分分析

主成分分析提供了降低数据复杂程度的数据压缩技术。PCA 方法是通过正交变换将一组可能存在相关关系的输入变量转换为不相关，转换后的组合即为主成分，PCA 集中分析所有方差来度量信息量的大小。另外的因子分析方法，则尝试识别相关性最大的因子，其只关注共享方差。这两种方法目标都是找出原始数据中信息集中的最重要因素。

5. 广义线性引擎模型

广义线性引擎(GLE)模型通过构建关联函数确认自变量和因变量的相关关系，该模型的优势是允许因变量非正态分布。它涵盖了广泛使用的统计模型，如用于正态分布响应的线性回归、用于二进制数据的逻辑模型、用于计数数据的对数线性模型、用于区间删失生存数据的互补重对数模型以及其他统计模型。

6. 广义线性模型

广义线性模型对一般线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。而且，该模型还允许因变量为非正态分布，它包括统计模型大部分的功能，其中包括线性回归、逻辑回归，用于计数数据的对数线性模型，以及区间删失生存模型。

7. 广义线性混合模型

广义线性混合模型(GLMM)扩展了线性模型，使得目标可以有非正态分布，通过指定的关联函数与因子和协变量线性相关，并且观测值可能相关。广义线性混合模型涵盖了从简单线性回归到复杂的非正态纵向数据多变量模型的各种模型。

8. COX 回归节点

COX 回归节点可以在已有的检查记录中建立时间事件的生存模型。该模型会生成一个生存函数，该函数可预测在给定时间(t)内对于所给定的输入变量值相关事件的发生概率。

4.2.9 聚类模型

聚类分析是一种把数据集划分成子集的过程。每一个子集构成一个簇,使簇之间的元素彼此相似,但簇与簇之间的元素又尽可能地彼此不相似。相异度是根据描述对象的属性值进行计算的,距离经常采用相异度度量方式。通常把一个簇内的对象作为一个整体对待。通常把聚类模型称为无监督模型,因为不存在用于判断分类结果的标准。

1. K-means 算法

K-means 算法将数据集聚类到不同分组。其中,相异度基于对象与簇中心的距离计算,与簇中心距离越近的对象可以划分为一簇。此算法的目标是每个对象与簇中心的距离平方和最小。通过不断迭代,进一步优化,直到聚类中心不再改变时,说明取到了最优解,即 K 的中心。

2. Kohonen 算法

Kohonen 算法是一种对数据集进行聚类的神经网络,将数据集中明显不同的类聚集到不同组中,训练完成后,相似点便已经聚集,异簇点远离,其优势是采用无监督学习的方式,对成员变量个数、资格等没有要求,也不需要指定目标字段。

3. 两步算法

两步算法使用两步聚类方法,第一步先对样本数据进行简单的处理,将原始数据放置到各个子簇类中,第二步使用层级聚类的方法,将子聚类逐步合并形成最大的簇类。

两步法的优点是:能够为训练数据自动估计最佳聚类,可以高效处理混合型的字段或者较大量数据集。不需要指定聚类的数量,通过检验多种聚类方法,然后取其中最有效的一种。还可以应用两步算法来检测离群值或其他异常值。

4.2.10 关联分析

关联是指在两个或多个变量之间存在某种规律性,但关联并不一定意味着因果关系。关联规则是在同一事件中出现的不同项目的相关性,关联分析是挖掘关联规则的过程。关联规则挖掘的核心是找出事务数据库中的所有强关联规则,其优点是对输入变量没有要求,缺点是运行效率较低。

1. Apriori 算法

Apriori 算法的思想是:先找出所有的频繁项集,然后由频繁项集产生强关联规则,这些规则必须满足最小支持度和最小置信度的要求。Apriori 算法的优点在于速度通常快一些。同时,Apriori 算法提供了 5 种不同的训练方法,应用更灵活。

2. 关联规则

关联规则节点与 Apriori 节点非常类似,但是存在一些明显的差异,其无法处理事务性数据,只能处理存储类型为列表的数据。可以与 IBM SPSS Analytic Server 配合处理大型数据。参数设置更多,支持较多个性化设置。在规则构建时,可以排除某些已经很明显的规则,减少资源浪费。如果输入字段为连续类型,可以使用离散化选项(分级化)进行分箱,输

出规则表时除了支持度和置信度外,还可以选择条件支持、部署能力、增益等,其中置信度作为默认规则排序标准,可在模型选项中对此进行更改,最大预测数表示最佳的预测规则数量,默认为3个。

3. CARMA 算法

CARMA 节点使用关联规则算法来发现数据中的关联规则。与 Apriori 算法相比,CARMA 节点不需要输入变量也能提取规则,在未选中事务格式时要求提供一个或多个内容字段,反之,要提供事务标识和一个名义字段。

4. 序列节点

序列模型也是关联分析的一种,它可以从连续的数据中发现模式规律,或者在面向时间的数据中发现模式。在模式挖掘过程中,其算法分为以下两步:首先发现常见序列,即频繁出现的序列,然后在线生成序列模式。序列节点要求指定一个编号字段或时间字段,至少一个内容字段,这些可以在字段选项卡上设置。由于序列模型采用的是 CARMA 算法,可以在“专家”选项卡中设置修剪值来调整修剪频率,以节省内存占用。另外,也可以设置内存中最大序列选项,以减少内存占用,但是这个数值要大于预期的结果序列数。

4.2.11 KNN 模型

最近邻(KNN)算法是一种用于分类和回归的监督学习方法,其采用向量空间模型来分类。

最近邻算法根据观测样本与其他样本在特征空间中的相似程度进行分类。在机器学习中,此方法不需要匹配原有模式或样例即可识别数据的模式。实现方法是:将靠近彼此的点视为相邻元素,当测试样本来了以后,计算其到每个观测值的距离。最近距离点的分类就是该测试样本的分类。

4.2.12 数据挖掘模式评估

评估节点的作用是评估和比较所使用的预测模型,让使用者选择最适合的模型,评估图表显示了模型对特定结果的预测优劣。根据观测值和预测值的置信度进行排序,将记录从高到低为每个分位数划分业务标准差。在散点图中,将以单独的线显示多个模型,从而进行比较。

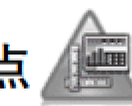
4.3 知识表示

数据可视化,指用图形或表格的形式显示信息。成功的可视化把数据及其信息转换成可视的形式,并且能够凸显出数据的特征,以及数据之间的关系和重要性。

使用可视化技术,可以帮助人们快速吸取大量信息,并观察到其中显著的模式和规律,让知识以最快的速度映射到人脑内。

4.3.1 图形节点

SPSS Modeler 中有导入图形和图表的功能，也有导出数据分布和它们关系的功能，这些图形化展示工具可以让分析人员对数据有更直观的理解和探查数据特征，不仅可以得到部分描述性统计结果，还有助于从中获得更多的分析思路和灵感。

1. 图形板节点 

图形板节点的作用是在单个节点上输出许多不同的图形（散点图、直方图、条形图等），以进行最佳选择。从第一个选项卡开始，选择所需数据字段，节点将提供一个适用于数据的图形类型的选项，节点自动过滤出适用于源的所有图形类型。在“详细”选项卡下，可以定义详细的选项或高级选项。

2. 散点图节点

散点图节点的作用是显示数值字段间的相互关系。在 SPSS Modeler 中，散点图可以开启 3D 效果图，单击 3D 按钮启用 Z 轴设置，在字段选择器中选择 Z 轴引用的字段。一旦图形生成之后，单击图形选项卡，单击 3D 按钮，即可将视图切换为 3D 图形。

3. 分布节点

分布节点的作用是显示不同维度衡量的出现次数，用于查看数据集中的程度。一般情况下，显示的是数据的分布状态、波动性和集中度。在数据分布不平衡度很高时，可以选用平衡节点来纠正不平衡度，然后使用分布节点。

4. 直方图节点

直方图节点的作用是显示不同字段的出现次数，可以有效地揭示不同阶段数据值的分布状态、波动性和集中度。在建模之前，常常用直方图检查数据，与分布节点类似，直方图常显示数据中的不平衡度。生成直方图时，还可以对横轴范围、分级、颜色、正态曲线、标题、标签、文字说明等选项进行编辑。

5. 多重散点图节点

多重散点图节点的作用是：在 X 变量上显示多个 Y 变量的关系图，用彩色线把 Y 变量连接起来。经常用于在一段时间内，多个变量随时间变化的效果图。

生成多重散点图时，还可以对交叠情形、标准化、交叠函数等选项进行设置。

6. 网络节点

网络节点的作用是说明两个或多个（分类）字段值之间关系的强度，连接线条显示关系链接，线的粗细表示关系强度，常用于挖掘频繁项集模式、关联分析和相关性分析。例如，购物篮中商品之间的关联，分析顾客的购物习惯。

4.3.2 数据输出

输出节点，在每个步骤中，提供了我们获取数据和模型信息的工具，可以导出、检查、分

析数据在每一步所处的状态和可能存在的问题。

1. 表节点

表节点的作用在于以表格的形式显示数据，当然，还可以写入到文件中，方便检查、导出用。在表节点中，常用设置“突出显示符合条件的记录”，通过输入 CLEM 表达式，针对指定的条件进行筛选，方便检查、导出所需数据。

2. 分析节点

分析节点的作用在于评估预测模型生成准确预测的能力，常常用于对一个或多个模型的预测值和实际值进行比较，可以比较模型间的优劣。

3. 数据审核节点

数据审核节点的作用在于全面检查数据，包括汇总统计量、直方图、分布和离群点、缺失值和极值的相关数据，将结果放到矩阵中，可用于排序，并生成图表和数据准备节点。

4. 变换节点

变换节点的作用是在不改变数据原始特征的情况下对数据进行变换，并将结果进行可视化，如果符合模型要求，则再应用于其他分析中。针对一些只适用于正态分布分析假设的分析方法，如回归、逻辑回归和判别分析等，原始数据不适用，常用的一种处理方法是对原始数据元素做变换，使其更接近正态分布，再进行回归分析。

4.3.3 数据导出

导出节点提供各种格式的导出数据的方式和工具，以便数据可以在各种形式下使用，包括文件格式、数据库及第三方数据分析软件。

1. 数据库导出节点

数据库导出节点是将数据写入与 ODBC 兼容的关系型数据库中。设置时，必须具备对应数据库写的权限。

2. 平面文件导出节点

平面文件导出节点的作用是将数据输出到已分隔的文本文件，可用于其他分析软件或电子表格分析，如导出以逗号或分号分隔的 csv 格式文本等。平面文件导出节点适用于导出其他分析或供软件读取电子表格用。

3. Excel 导出节点

Excel 导出节点的作用是输出 Excel 文件，可以指定字段、Excel 中的起始单元格，选择是建立新文件，还是插入到原有文件中等，还可以等导出成功后直接打开 Excel 文件。

第 5 章

香水销售分析

法国著名的诗人保罗·瓦莱利曾说："不擦香水的女人没有未来。"香水对于现代都市女性，不仅是生活品位的标志，更是个人气质的象征。对于男士来说，使用香水也是提升个人魅力的途径。随着经济发展以及人民生活水平的提高，国内消费者对香水产品的消费需求快速增长，曾经作为奢侈品的香水，逐渐成为人们的日常生活用品。

我国的香水行业较欧美国家起步晚，所占市场份额小，目前尚处于成长期。这同时也说明了我国的香水市场有巨大的潜力。许多国际大牌香水制造商正在努力提升它们在中国市场的份额，竞争未来的巨大市场前景。全球范围内，香水市场是一个市值 438.9 亿美元的产业，每年至少有 300 种新品种上市。

根据中国产业信息网的统计，2015 年我国香水市场规模为 185 亿元，环比增长 15.6%。我国香水市场的快速发展也带来了不同香水产品的大量涌现，那么到底什么样的香水产品销量更好，更受消费者欢迎呢？这些香水产品又有什么样的特点呢？本章以从某电商网站上抓取到的香水产品销量数据分析香水销售的影响因素，为香水销售商判定采购计划以及用户选择香水提供依据。

5.1 香水销售数据预处理

本案例从某电商网站抓取了 1009 条香水产品销售数据，包含了香水产品的商品名称、产品毛重、商品产地、包装、香调、净含量、分类、适用性别、适用场所、价格，以及评价数。

"评价"字段的数据包含混合的中文和数字，末尾有一个"+"号。"+"号很容易通过 Excel 替换成" "(空字符串)的方法除去，但是将"万"转换成准确的数值结果，采用 Excel 或者其他现成的工具，并不容易实现。因此，采用 Python 编程处理"评价"和"适用场合"字段，将评价数量转换成数值。

"适用场所"字段包含多个场所，如果要拆分成多个字段，首先要算出所有记录的场所合集，这个步骤也不容易通过现有工具实现。

这两个预处理步骤，最终采用 Python 编程实现。通过 Python 脚本生成了新的字段。将商品产地中的"中国大陆""广东""浙江义乌"等统一替换成"中国"。"适用场所"字段分解成"旅行""其他""约会""情趣"等 8 个字段，其类型是 0、1 类型，若该香水产品有对应的适用场所，则设置为 1，否则设置为 0。例如，第一条冰希黎的香水数据，使用场所为旅行、约会、情趣、商务、party 聚会。将类似"1.9 万＋"格式的"评价"字段的值转换为"19 000"，其类型是数值类型。

Python 程序没有替换原有字段，而是生成一个新字段。其优点是，不会丢失原始数据，而且可以通过肉眼复查，检查是否有预处理出错的情况。经过人工审查，经过 Python 程序预处理的数据符合原数据。图 5.1 是 Python 预处理完成的香水数据。

1	商品名称	商品产地	包装	香调	净含量	分类	性别	适用场所	价格	评价	旅行	其它	约会	情趣	商务	日常	party聚会	运动
2	冰希黎695600860	中国	Q版香水	花果香调	1ml-15ml	浓香水EDP	女	日常，约会，情趣，商务，party聚会，旅行	10	19000	1	0	1	1	1	1	1	0
3	冰希黎695600860	中国	Q版香水	混合香调	1ml-15ml	浓香水EDP	女	日常，约会，party聚会，运动，旅行	10	19000	1	0	1	0	0	1	1	1
4	（免邮）上海老[		独立装	花果香调	31ml-100ml	浓香水EDP	女	日常，约会，商务，party聚会，旅行	18	90	1	0	1	0	1	1	1	0
5	法颂浪漫梦境女士		Q版香水	花果香调	1ml-15ml	固体香水/香膏	女	日常，约会，商务，party聚会，运动，旅行	22	30	1	0	1	0	1	1	1	1
6	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，旅行	24	100	1	0	1	0	1	1	1	1
7	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会，商务，旅行	24	100	1	0	0	0	0	0	1	0
8	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会，商务，party聚会，运动，旅行	24	39000	1	0	0	0	0	1	0	0
9	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	party聚会，旅行	24	39000	1	0	1	0	1	1	0	0
10	美颖香水女士桂花		独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会，旅行	25	50	1	0	1	0	0	1	0	0
11	冰希黎695600860	中国	独立装	混合香调	其它	固体香水/香膏	女	日常，约会，商务，party聚会，运动，旅行	25	19000	1	0	1	0	1	1	1	1
12	冰希黎695600860	中国	独立装	混合香调	其它	固体香水/香膏	女	日常，约会，情趣，商务，party聚会，旅行	25	19000	1	0	1	1	1	1	1	0
13	艾诗止汗香体走珠		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会，运动，旅行	26	300	1	0	1	0	0	1	0	1
14	艾诗止汗香体走珠		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会，运动，旅行	26	300	1	0	1	0	0	1	0	1
15	阿迪达斯女士净h	中国		海洋香调	31ml-100ml	香体走珠	女		27	2700	0	1	0	0	0	1	0	1
16	阿迪达斯香体液	中国	独立装	混合香调	31ml-100ml	淡香水EDT	女	日常，运动，其它	27	2900	1	1	1	1	1	1	1	1
17	阿迪达斯香体液	中国		混合香调	31ml-100ml	香体走珠	女	日常，约会，情趣，商务，party聚会，运动，旅行，;	27	5900	0	0	0	0	0	0	0	0
18	妮维雅走珠止汗		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会，party聚会，运动	29	1300	0	0	1	0	0	1	1	1
19	Chanel香奈儿邂逅		独立装	花果香调	31ml-100ml	淡香水EDT	女	约会，情趣，商务，party聚会	29	40	1	0	1	0	0	1	0	0
20	美颖香水女士桂花	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会，旅行	29	300	0	0	0	0	0	1	0	0

图 5.1 Python 预处理完成的香水数据

对香水产品的价格和评价数进行离散化处理，将价格等间距分为 6 个等级，记为低、较低、中等、较高、高、非常高，对应价格区间分别为(0,100]，(100,300]，(300,500]，(500,700]，(700,1000]，1000 以上；同样，将评价数等间距分为 7 个等级，记为非常低、低、较低、中等、较高、高、非常高，对应价格区间分别为(0,100]，(100,500]，(500,1000]，(1000,2000]，(2000,5000]，(5000,10 000]，10 000 以上。

将价格和评价数离散化后的变量记为"价格等级"和"销量等级"，在 SPSS Modeler 18.0 中使用导出节点进行处理，如图 5.2 和图 5.3 所示。

公式:

```
if (价格 <= 100) then '低'
 else if (价格 <= 300) then '较低'
  else if (价格 <= 500) then '中等'
   else if (价格 <= 700) then '较高'
    else if (价格 <= 1000) then '高'
     else '非常高'
    endif
   endif
  endif
 endif
endif
```

图 5.2 "价格等级"导出公式

公式:

```
if (评价 <= 100) then '非常低'
 else if (评价 <= 500) then '低'
  else if (评价 <= 1000) then '较低'
   else if (评价 <= 2000) then '中等'
    else if (评价 <= 5000) then '较高'
     else if (评价 <= 10000) then '高'
      else '非常高'
     endif
    endif
   endif
  endif
 endif
endif
```

图 5.3　“销量等级”导出公式

对香水产品的适用场合进行数量统计，得到新字段“适用场合数量”。图 5.4 显示了最终处理得到的香水产品数据。

表格　注解

	商品名称	商品产地	包装	香调	净含量	分类	性别	适用场所	价格	评价	旅行	其他	约会	情趣	商务	日常	party...	运动	适用场合数量	价格等级	销量等级
1	冰希黎6...	中国	Q版香水	花果香调	1ml-15ml	浓香水EDP	女	日常，约会...	9.9...	190...	1.0...	0.0...	1.0...	1.0...	1.0...	1.0...	1.000	0.0...	6.000	低	非常高
2	冰希黎6...	中国	Q版香水	混合香调	1ml-15ml	浓香水EDP	女	日常，约会...	9.9...	190...	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	1.000	1.0...	5.000	低	非常高
3	（免邮...		独立装	花果香调	31ml-100ml	浓香水EDP	女	日常，约会...	18....	90.0...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	0.0...	5.000	低	非常低
4	法颂浪...		Q版香水	花果香调	1ml-15ml	固体香水/...	女	日常，约会...	22....	30.0...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	1.0...	6.000	低	非常低
5	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，旅行	23....	100....	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	1.0...	6.000	低	非常低
6	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会...	23....	100....	1.0...	0.0...	0.0...	0.0...	0.0...	0.0...	1.000	0.0...	2.000	低	非常低
7	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	日常，约会...	23....	390...	1.0...	0.0...	0.0...	0.0...	0.0...	1.0...	0.000	0.0...	2.000	低	非常高
8	雅芳香水	中国	独立装	东方香调	1ml-15ml	淡香水EDT	女	party聚会...	23....	390...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	0.000	0.0...	4.000	低	非常高
9	美顿香...		独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	25....	50.0...	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	0.0...	3.000	低	非常低
10	冰希黎6...	中国	独立装	混合香调	其它	固体香水/...	女	日常，约会...	25....	190...	1.0...	0.0...	1.0...	0.0...	1.0...	1.0...	1.000	1.0...	6.000	低	非常高
11	冰希黎6...	中国	独立装	混合香调	其它	固体香水/...	女	日常，约会...	25....	190...	1.0...	0.0...	1.0...	1.0...	1.0...	1.0...	1.000	0.0...	6.000	低	非常高
12	艾诗止...		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会...	25....	300....	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	1.0...	4.000	低	低
13	艾诗止...		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会...	25....	300....	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	1.0...	4.000	低	低
14	阿迪达...	中国		海洋香调	31ml-100ml	香体走珠	女		26....	270...	0.0...	1.0...	0.0...	0.0...	0.0...	1.0...	0.000	1.0...	3.000	低	较高
15	阿迪达...	中国	独立装	混合香调	31ml-100ml	淡香水EDT	女	日常，运动...	26....	290...	1.0...	1.0...	1.0...	1.0...	1.0...	1.0...	1.000	1.0...	8.000	低	较高
16	阿迪达...	中国		混合香调	31ml-100ml	香体走珠	女	日常，约会...	26....	590...	0.0...	0.0...	0.0...	0.0...	0.0...	0.0...	0.000	0.0...	0.000	低	高
17	妮维雅...		独立装	花果香调	31ml-100ml	香体走珠	女	日常，约会...	28....	130...	0.0...	0.0...	1.0...	0.0...	0.0...	1.0...	1.000	1.0...	4.000	低	中等
18	Chanel...		独立装	花果香调	31ml-100ml	淡香水EDT	女	约会，情趣...	29....	40.0...	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	0.0...	3.000	低	非常低
19	美顿香...	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	29....	300....	0.0...	0.0...	0.0...	0.0...	0.0...	1.0...	0.000	0.0...	1.000	低	低
20	美顿香...	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	29....	300....	0.0...	0.0...	1.0...	1.0...	1.0...	0.0...	1.000	0.0...	4.000	低	低
21	美顿香...	中国	独立装	花果香调	31ml-100ml	淡香水EDT	女	日常，约会...	29....	300....	1.0...	0.0...	1.0...	0.0...	0.0...	1.0...	0.000	0.0...	3.000	低	低

图 5.4　最终处理得到的香水产品数据

5.2　香水销售数据统计分析

对 1009 条香水产品的价格进行描述分析，约 92.47% 的产品价格在 900 元以下，如图 5.5 所示。最大值为 2212 元，在样本集中可查询到对应产品为香奈儿机会/机遇/黄色邂逅女士香水 50/100mL/持久淡香精 EDP EDP100mL。

产品的评论数在一定程度上代表了产品的销量，因此用评论数来代替产品的销量。对所有产品的销量进行统计分析，结果如图 5.6 所示，香水产品的销量两极分化明显，有 58.87% 的产品销量不足 1000，有约 10.0% 的产品销量大于 10 000，其中最大值为 100 000+，在样本数据集中对应的产品为菲拉格慕(Ferragamo)梦中情人女士香水礼盒(香水 100mL+身体乳 150mL)。

将预处理完成的 Excel 数据导入 SPSS。发现“商品产地”存在大量空值。于是，在分析商品产地分布之前，使用 SPSS 的“记录选项”→“选择”组件对数据进行过滤。过滤规则是【商品产地＝" "】。过滤后，数据记录数目减少至 489 条。获取的香水产品产地分布情况如

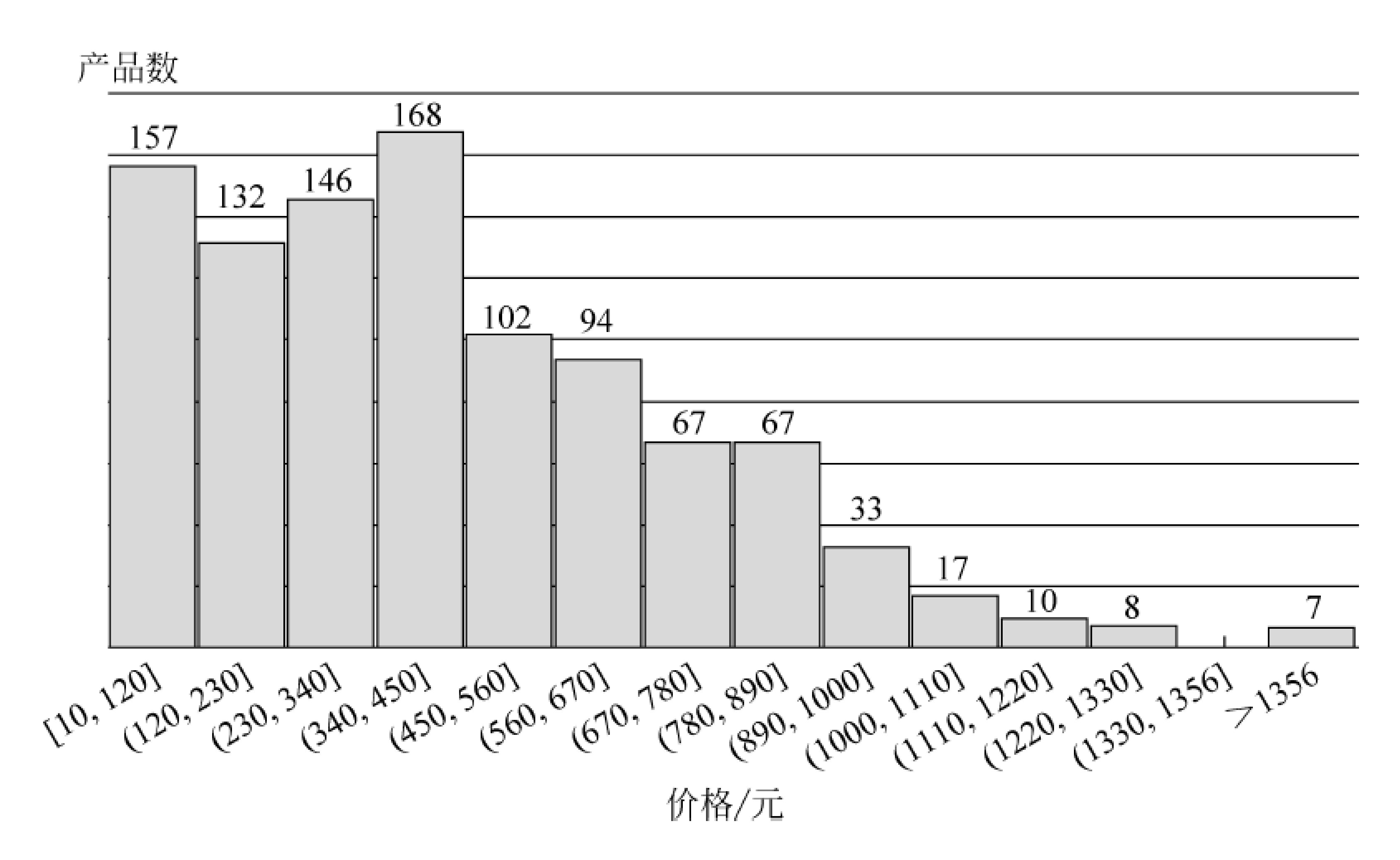


图 5.5 香水产品价格描述分析图

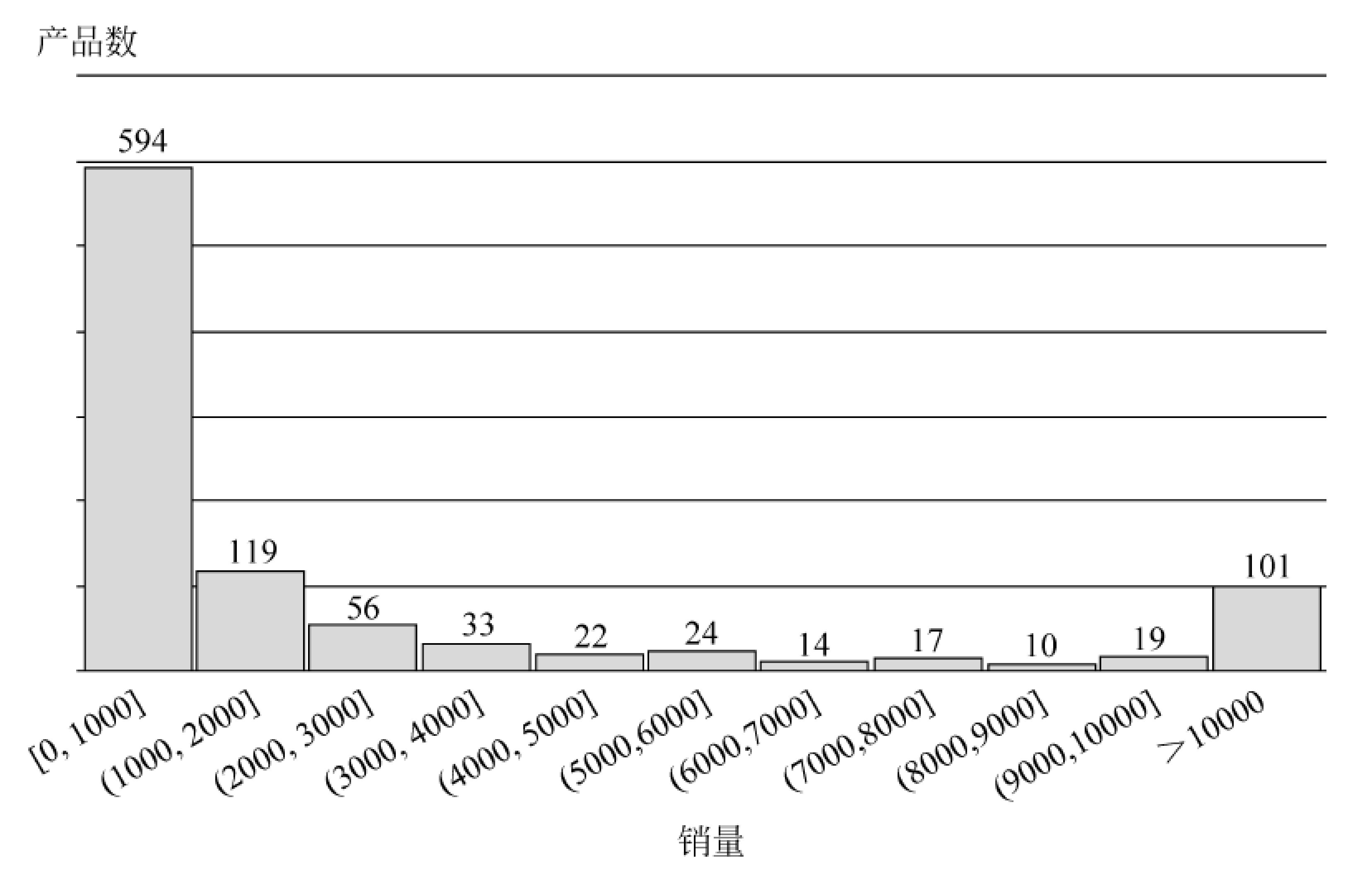


图 5.6 香水产品销量描述分析图

图 5.7 所示，所有香水产品的产地中，法国占据了绝对比例，为 47.99%。德国和西班牙产的香水种类较少，分别为 3.82%和 3.01%，如图 5.7 所示。

由于商品的评价数跨度比较大，且商品销量的两极分化严重，如果直接用评价数来绘制箱型图，会产生大量的离群点，不够直观。因此，对商品的评价数以 2 为底求对数值，再按照各个字段对“评价对数值”绘制箱型图。

用箱型图描述各产地香水的销量分布，如图 5.8 所示。从图 5.8 中可以看出，与其他国家的香水产品相比，西班牙和英国的香水产品销量明显偏低，而德国、法国、美国、意大利和中国的香水产品则没有明显差别。

图 5.9 显示了各包装香水销量箱型图。从图 5.9 中可以看出，组合装香水的整体销量要高于其他包装的香水，因为组合装的香水价格往往比较优惠，对价格敏感的消费者有较大

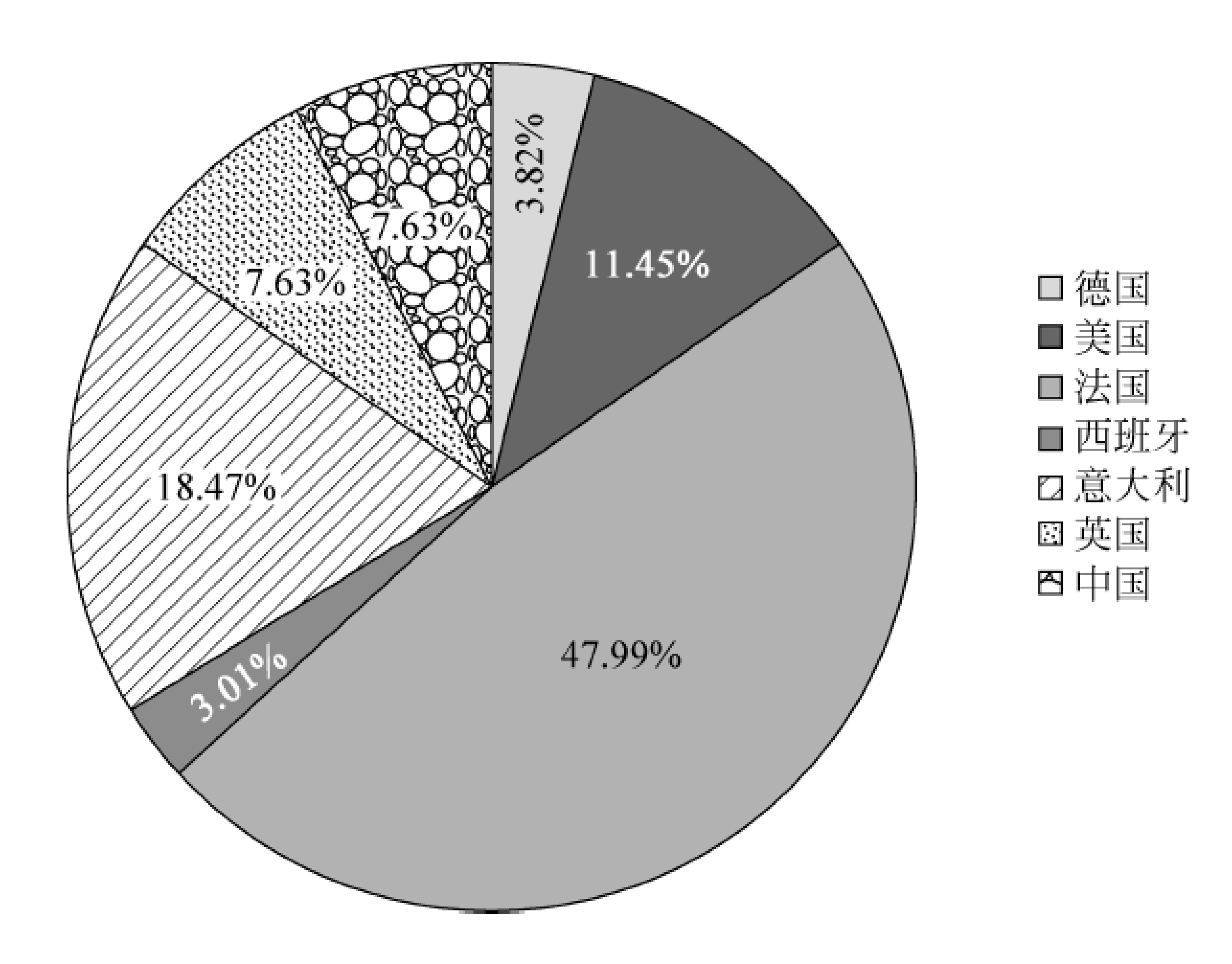


图 5.7　香水产品产地分布图

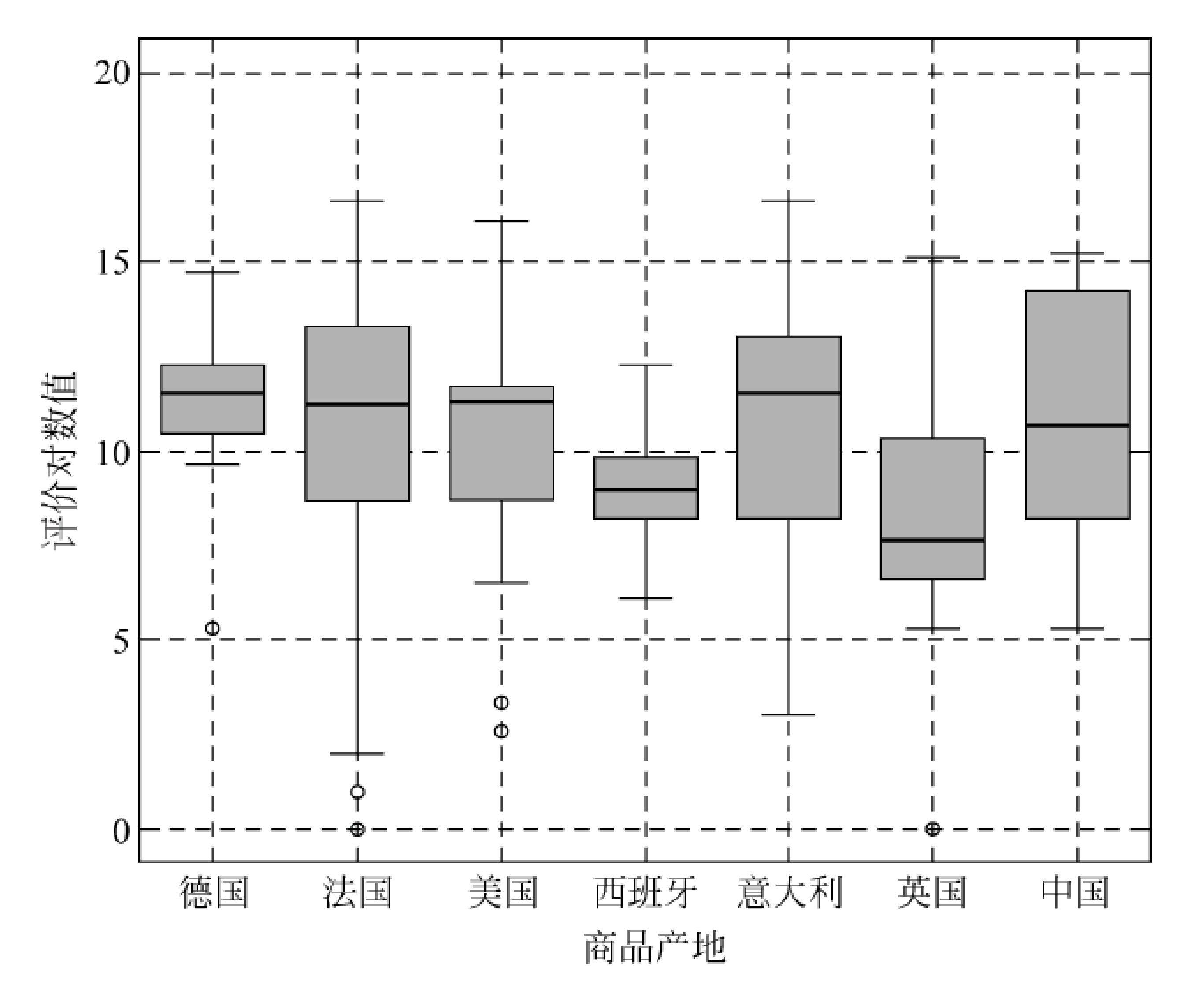


图 5.8　各产地香水销量箱型图

的吸引力。而限量版香水的销量要明显低于其他包装的香水，主要原因是由于限量版香水的发行量少且价格较高。而Q版香水、独立装香水、礼品套装和其他包装的香水，销量则没有明显的差别。

将不同香调的香水销量绘制箱型图，如图 5.10 所示。可以看出，花果香调和混合香调的香水产品整体销量要略高于其他香调的香水，而东方香调和其他香调的香水整体销量偏低。海洋香调和木质香调的香水销量介于两者之间。东方女性使用香水的习惯较西方女性保守，偏好轻盈简单的清淡味道，因此花果香调的香水卖得最好。木质香调等较浓郁的香水遮盖体味功能较强，比较适合西方人，在以年轻女性为主力消费者的中国市场表现一般。

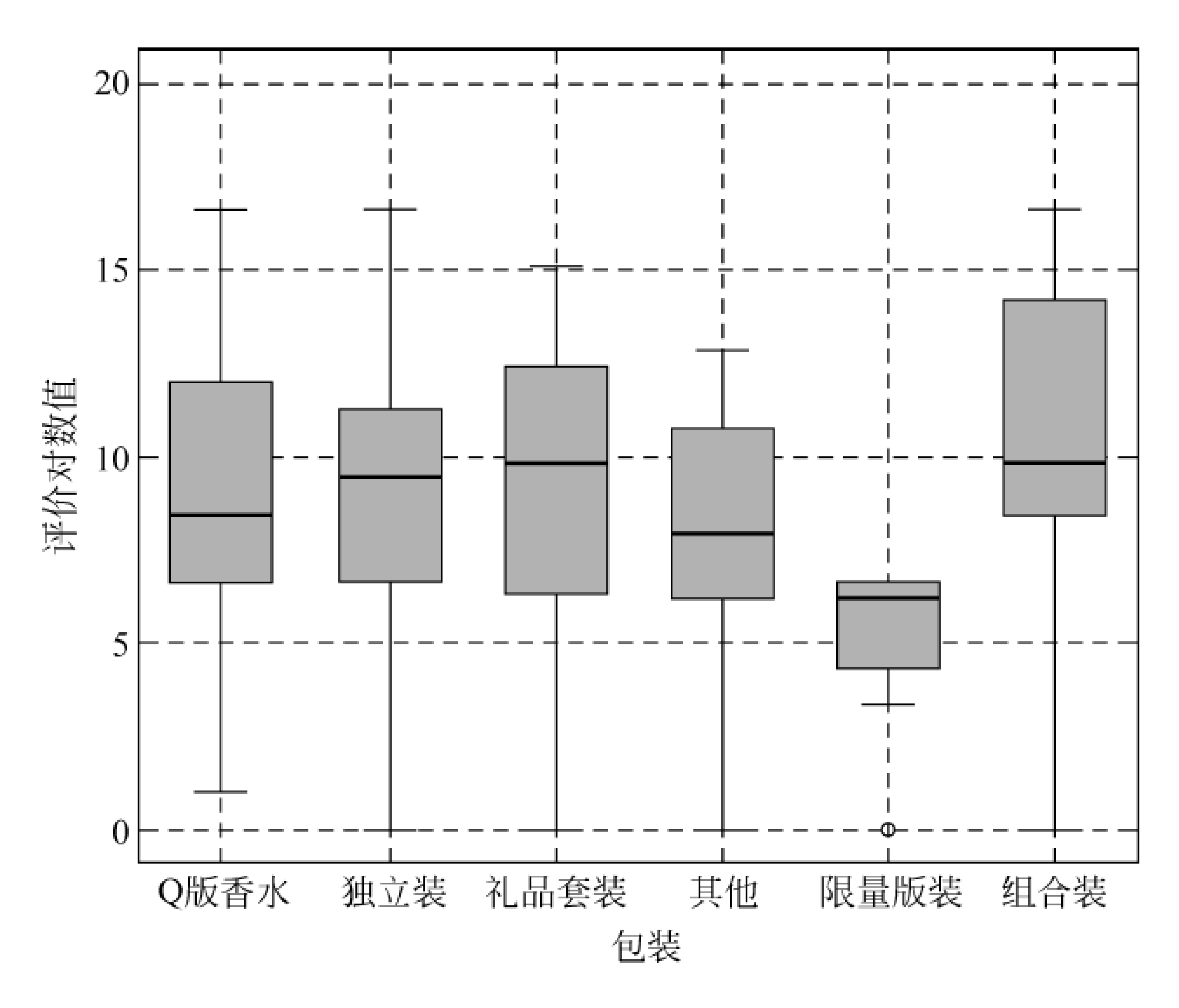


图 5.9 各包装香水销量箱型图

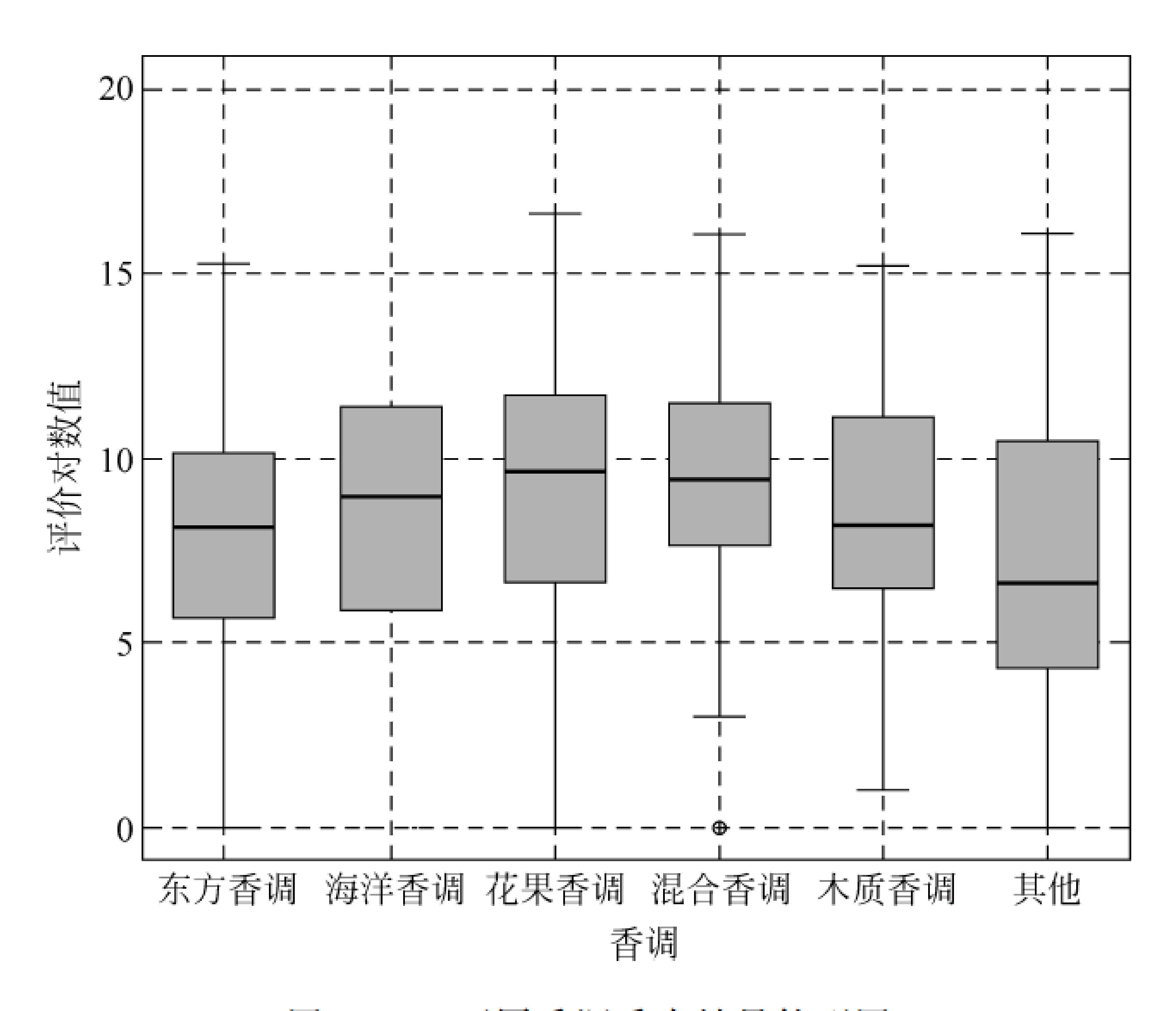


图 5.10 不同香调香水销量箱型图

如图 5.11 所示，在净含量方面，包装较小的产品销量比较高，包括 1～15mL，16～30mL，31～100mL。包装小的香水产品便携性强，而且我国大部分的香水使用者使用需求并不如欧美国家的消费者高，因此偏好小包装的香水。而 101～200mL 以及 200mL 以上规格的香水，不方便携带，而且如果不及时使用完毕，会有变质等问题，因此大规格的香水销量比小包装的香水要低。

按分类分析前，因为除了淡香水 EDT 和浓香水 EDP 外的其他种类，如香体走珠、固体香水/香膏、发香雾等类别的样本个数较少，所以统称为“其他”分类，如图 5.12 所示。不同分类的香水方面，淡香水 EDT 和浓香水 EDP 的销量好。淡香水 EDT 味道清淡，符合东方

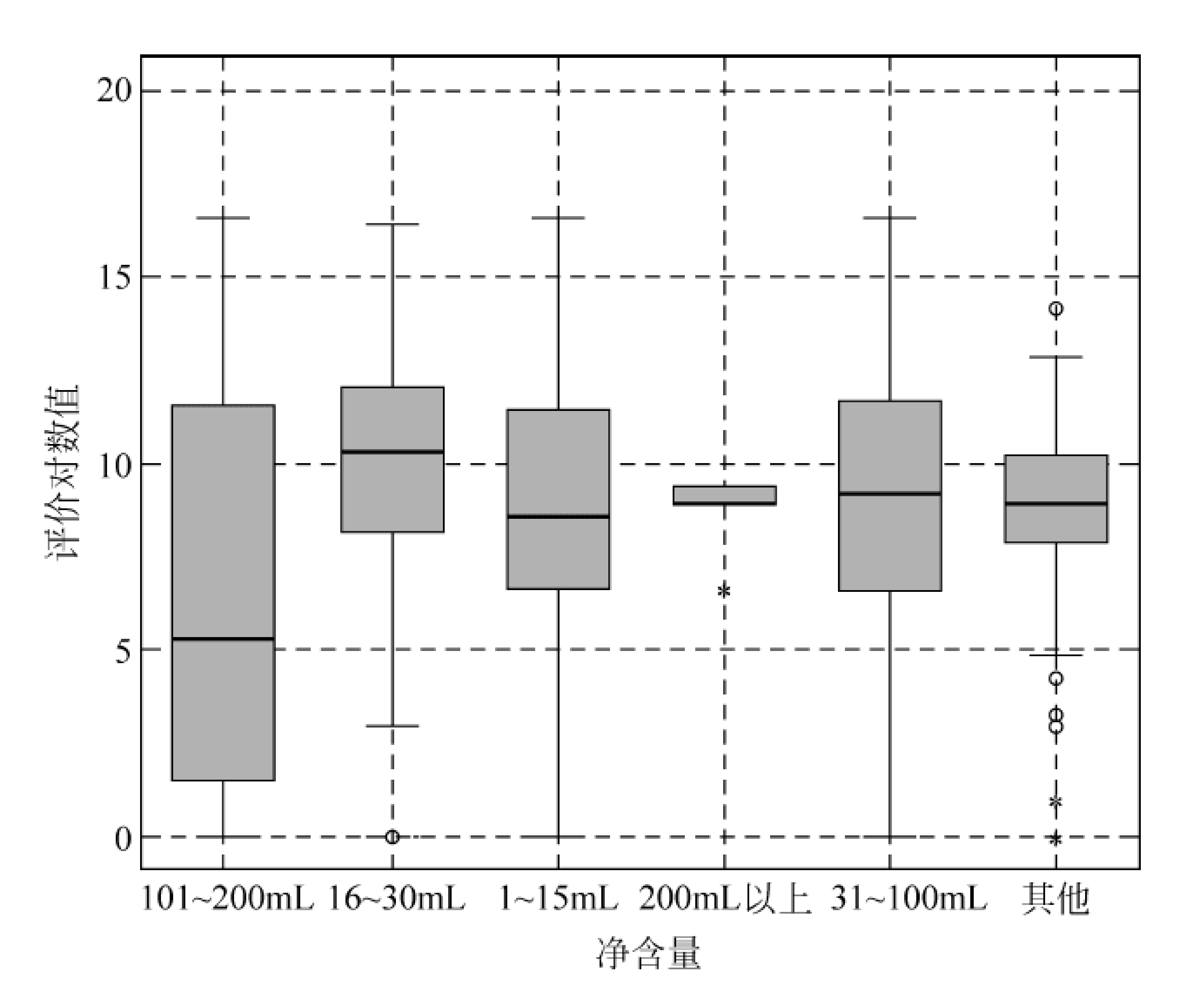


图 5.11 不同净含量香水销量箱型图

女性的消费特征。浓香水 EDP 主要针对年纪较大的商务女性和中年女性，也有一定的市场。其他类别的香水整体销量要低于淡香水 EDT 和浓香水 EDP。

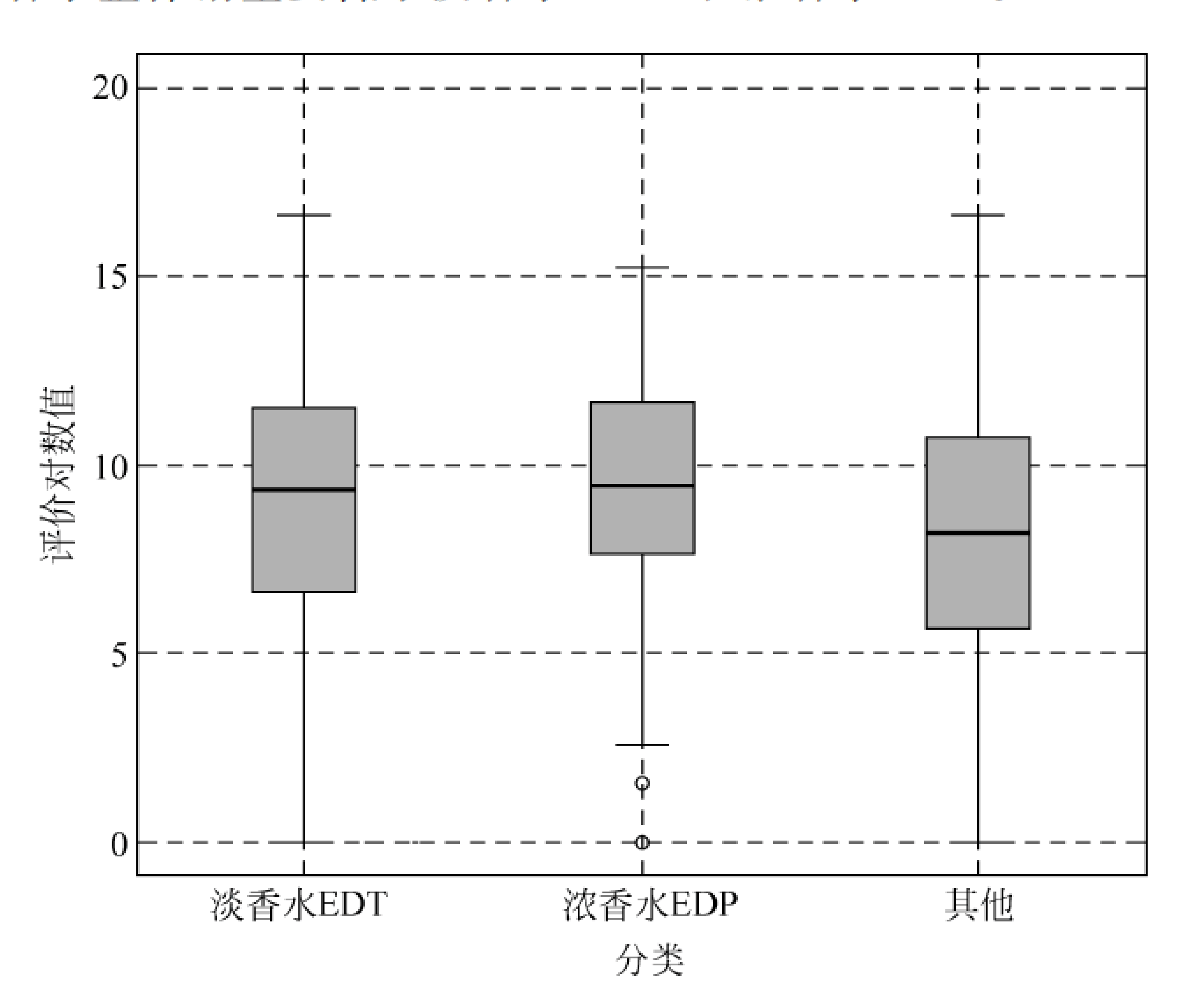


图 5.12 各分类香水销量箱型图

按照 1/3、2/3 分位数划分价格为高、中、低价位，对各使用场合、不同价位香水进行平均销量统计，如图 5.13 所示。其中，适用于 party 聚会、旅行、约会、商务、运动的香水比较受欢迎。情趣、日常和其他适用场合的香水销量明显低于其他场合。

在价格敏感性方面，所有使用场合的香水产品都体现了明显的价格敏感性。如图 5.14 所示，低价位、中价位、高价位的香水销量依次降低。其中，情趣、商务的产品对价格最敏感。

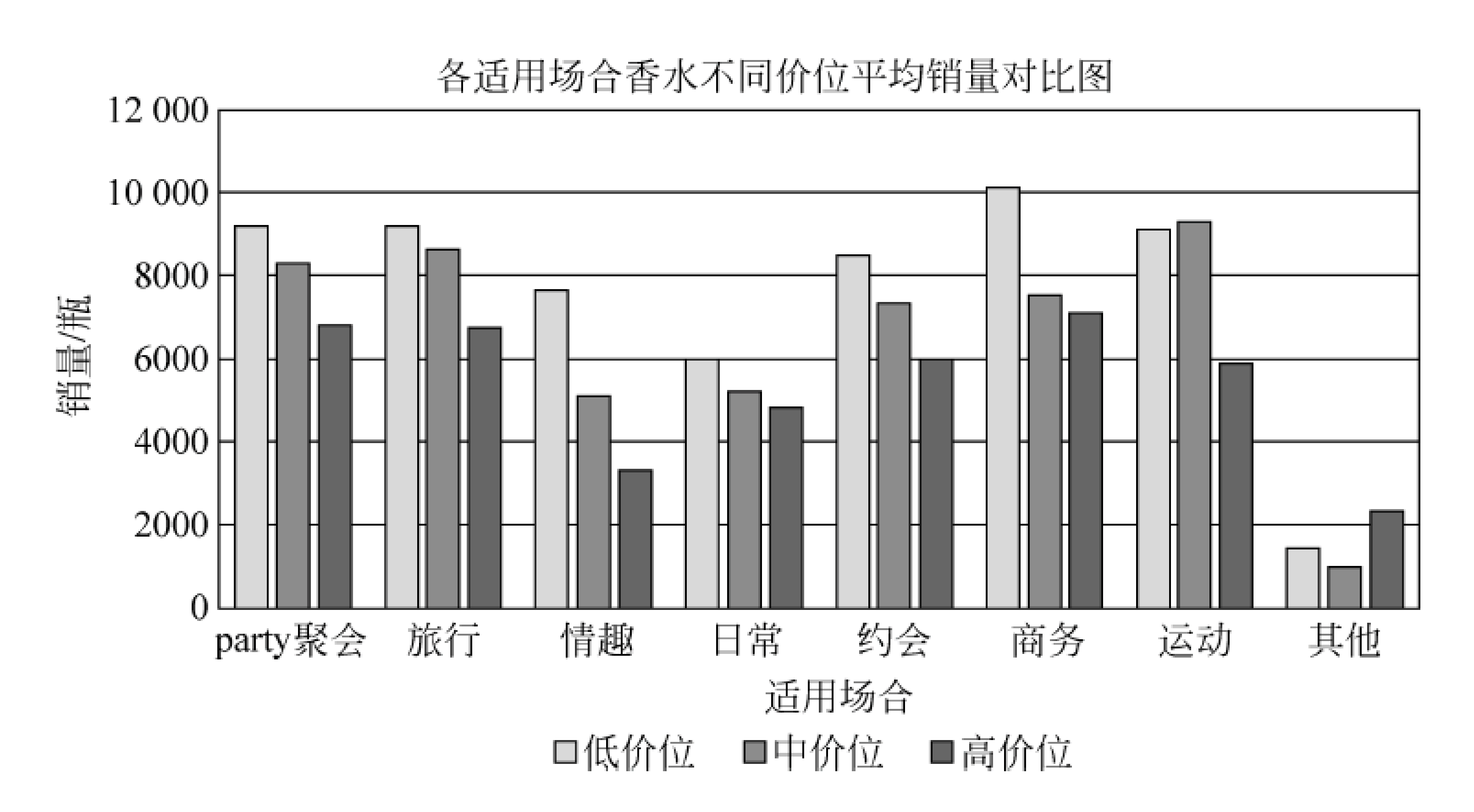


图 5.13 各适用场合香水不同价位平均销量(评论条数)对比图

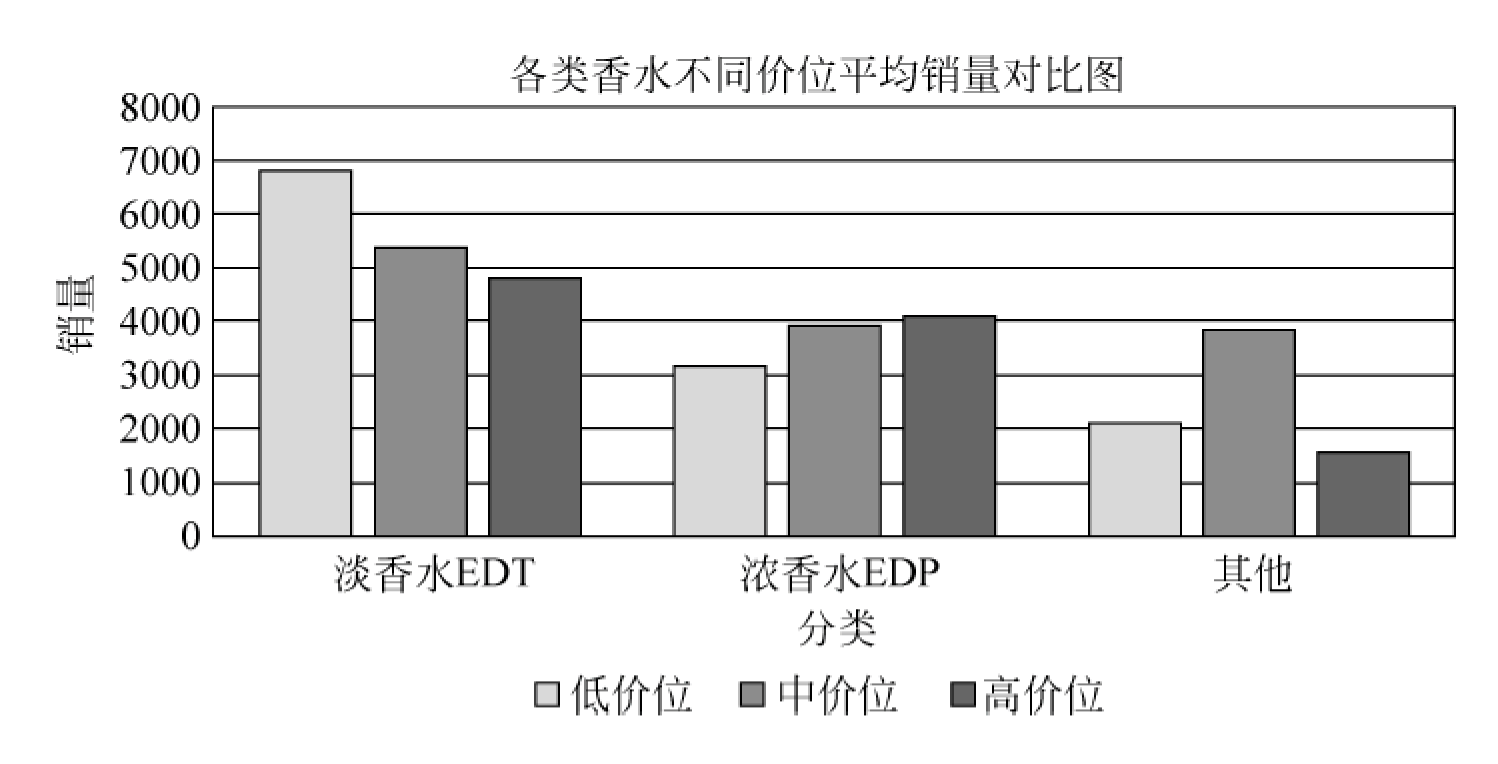


图 5.14 各类香水不同价位平均销量对比图

对不同种类的香水按照价位进行平均销量统计,可以发现淡香水 EDT 对价格敏感,价格低的产品销量好;而由于浓香水 EDP 主要的消费群体为年龄较大的中年女性和商务女性,对价格不敏感,价格越高的产品反而销量越高。而其他类别的香水,中价位的产品销量最好。

5.3 影响香水销量的因素分析

将 Python 预处理完成的 Excel 数据导入 SPSS。发现“商品产地”和“包装”存在大量空值。如果不进行处理,那么在分析影响销量的因素时使用 SPSS 的“记录选项”→“选择”组件,对数据进行过滤。过滤规则是[商品产地=" " or 包装=" "],如图 5.15 所示。过滤后,数据记录数目减少至 487 条。

使用“过滤器”节点,过滤掉本次分析不需要的字段,选择恰当的字段挖掘影响销量等级的因素。本次因变量为“销量等级”,自变量为“商品产地”“包装”“香调”“净含量”“分类”“性别”“适用场合数量”和“价格等级”。“过滤器”节点设置如图 5.16 所示。

使用“类型”节点,将“销量等级”字段设置为目标,其他字段设置为输入,如图 5.17 所示。使用 C5.0 决策树算法,挖掘影响香水产品销量等级的因素,图 5.18 展示了 SPSS Modeler 18.0 中的处理流程。



图 5.15　去除含空值记录

过滤器　注解

字段：已输入 21 个，已过滤 12 个，已重命名 0 个，已输出 9 个

字段	过滤器	字段
商品名称	×→	商品名称
商品产地	→	商品产地
包装	→	包装
香调	→	香调
净含量	→	净含量
分类	→	分类
性别	→	性别
适用场所	×→	适用场所
价格	×→	价格
评价	×→	评价
旅行	×→	旅行
其他	×→	其他
约会	×→	约会
情趣	×→	情趣
商务	×→	商务
日常	×→	日常
party聚会	×→	party聚会
运动	×→	运动
适用场合数量	→	适用场合数量
价格等级	→	价格等级
销量等级	→	销量等级

图 5.16　过滤掉不需要的字段

类型　格式　注解

读取值　清除值　清除所有值

字段	测量	值	缺失	检查	角色
商品产地	名义	"",中国,德...		无	输入
包装	名义	"",Q版香水,...		无	输入
香调	名义	"",东方香调...		无	输入
净含量	名义	"","101ml-...		无	输入
分类	名义	"",其他,发...		无	输入
性别	名义	女,通用		无	输入
适用场合数量	连续	[0.0,8.0]		无	输入
价格等级	名义	中等,低,较...		无	输入
销量等级	名义	中等,低,较...		无	目标

图 5.17　类型设置示意图

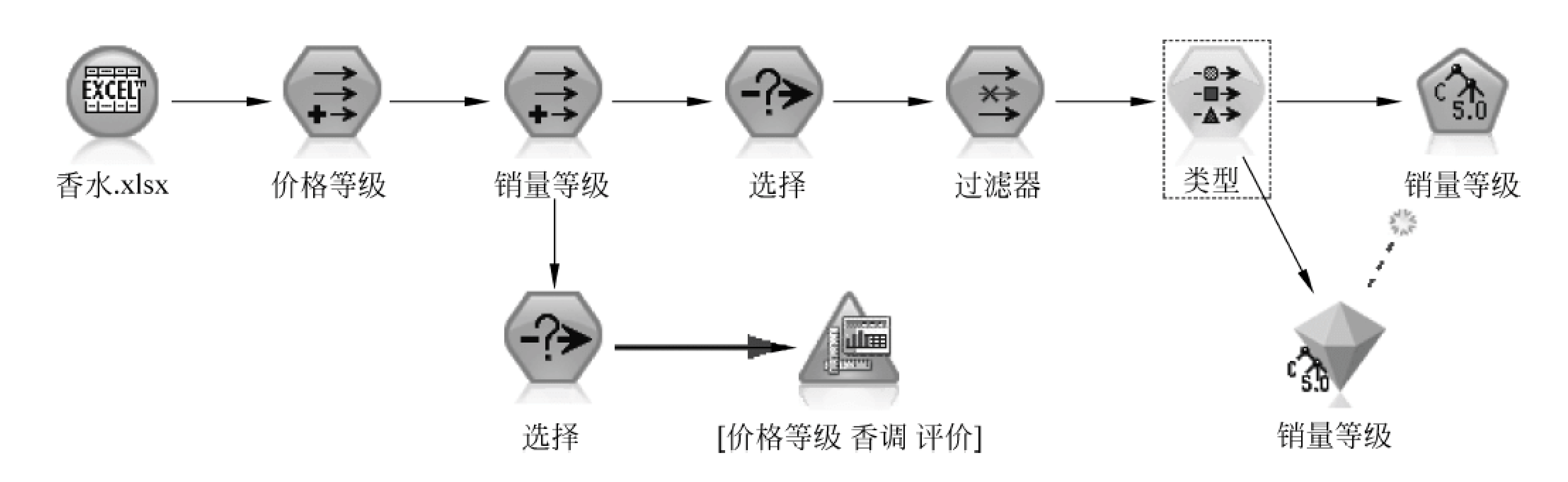


图 5.18 决策树具体构造流程

预测变量重要性如图 5.19 所示。在影响香水产品销量的因素中，商品产地是最重要的，其次是包装、适用场合数量和香调，它们对销量有较大的影响。净含量、性别、价格等级、分类对销量的影响较小。

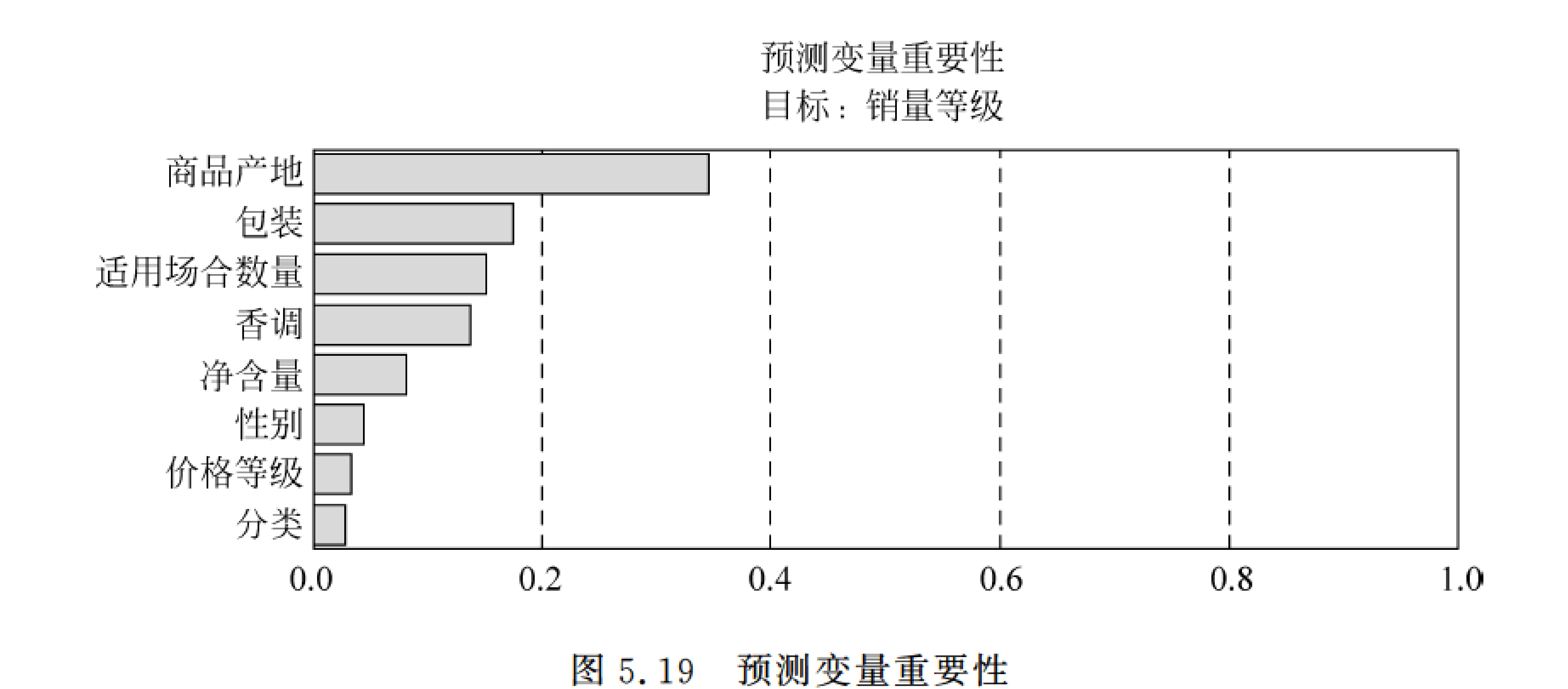


图 5.19 预测变量重要性

图 5.20 显示了具体生成的决策树。可以得到如下结论：

(1) 中国和法国生产的香水最受消费者欢迎，整体销量等级为“非常高”。法国香水有着悠久的历史，法国是世界上最著名的香水产地，拥有大量的客户，因此整体销量非常高。而中国香水比较符合东方人的口味，而且价格较低，能够吸引大量的消费者。

(2) 在中国和法国生产的香水中，消费者更加看重的是香水的香调。整体销量最高的“花果香调”在国产香水中销量反而较低；整体销量较低的“木质香调”销量却较高。说明国产香水中“花果香调”的香水产品不受消费者欢迎，应当适当调整香水的生产、销售策略吸引更多消费者。

(3) 德国、意大利和美国的香水整体销量较高。但是第二层中，对于德国香水，消费者更加注重的是香水的净含量；对于意大利香水，消费者更加看重价格；对于美国香水，消费者更加看重包装。

(4) 英国和西班牙的产品销量较低。对于英国香水，消费者更加看重香调。

```
商品产地 = [模式: 非常高] ⇨ 非常高
⊟ 商品产地 = 中国 [模式: 非常高]
    香调 in ["" "木质香调" "海洋香调"] [模式: 非常高] ⇨ 非常高
  ⊞ 香调 in ["东方香调"] [模式: 非常高]
    香调 in ["其他"] [模式: 非常低] ⇨ 非常低
  ⊞ 香调 in ["混合香调"] [模式: 非常高]
  ⊞ 香调 in ["花果香调"] [模式: 低]
⊟ 商品产地 = 德国 [模式: 较高]
    净含量 in ["" "101ml-200ml" "200ml以上"] [模式: 较高] ⇨ 较高
    净含量 in ["16ml-30ml"] [模式: 较高] ⇨ 较高
    净含量 in ["1ml-15ml"] [模式: 非常高] ⇨ 非常高
  ⊞ 净含量 in ["31ml-100ml"] [模式: 较高]
    净含量 in ["其他"] [模式: 较低] ⇨ 较低
⊟ 商品产地 = 意大利 [模式: 较高]
  ⊞ 价格等级 = 中等 [模式: 较高]
  ⊞ 价格等级 = 低 [模式: 较高]
  ⊞ 价格等级 = 较低 [模式: 非常高]
  ⊞ 价格等级 = 较高 [模式: 较高]
    价格等级 = 非常高 [模式: 较高] ⇨ 较高
  ⊞ 价格等级 = 高 [模式: 低]
⊟ 商品产地 = 法国 [模式: 非常高]
    香调 in [""] [模式: 非常高] ⇨ 非常高
    香调 in ["东方香调" "海洋香调"] [模式: 非常低] ⇨ 非常低
  ⊞ 香调 in ["其他"] [模式: 非常低]
  ⊞ 香调 in ["木质香调"] [模式: 较低]
  ⊞ 香调 in ["混合香调"] [模式: 较高]
  ⊞ 香调 in ["花果香调"] [模式: 非常高]
⊟ 商品产地 = 美国 [模式: 较高]
    包装 in ["" "其他" "限量版装"] [模式: 较高] ⇨ 较高
    包装 in ["Q版香水"] [模式: 非常高] ⇨ 非常高
  ⊞ 包装 in ["独立装"] [模式: 较高]
    包装 in ["礼品套装"] [模式: 低] ⇨ 低
    包装 in ["组合装"] [模式: 较低] ⇨ 较低
⊟ 商品产地 = 英国 [模式: 非常低]
    香调 in ["" "东方香调" "其他"] [模式: 非常低] ⇨ 非常低
    香调 in ["木质香调"] [模式: 低] ⇨ 低
    香调 in ["海洋香调"] [模式: 非常低] ⇨ 非常低
  ⊞ 香调 in ["混合香调"] [模式: 非常低]
  ⊞ 香调 in ["花果香调"] [模式: 中等]
商品产地 = 西班牙 [模式: 低] ⇨ 低
```

图 5.20　销量影响因素决策树分析结果

5.4　香水适用场所关联分析

对香水适用场所进行关联分析。对源数据进行预处理，将适用场所分隔开，生成不同的字段，总共为 8 类。将含有该类适用场所的值设置为 1.0，否则设置为 0.0。在关联分析前过滤掉除适用场所以外的所有本次分析不需要的字段，将所有适用场所的类型设置为任意，如图 5.21 和图 5.22 所示。

进行关联分析时，采用 Apriori 算法，将最低条件支持度设置为 55%，最小规则置信度设置为 90%，运行 Apriori 节点，最终得到 12 条关联规则。IBM SPSS Modeler 18.0 中具体的操作流程如图 5.23 所示。

图 5.24 显示了对香水产品适用场所进行关联分析后的结果。可以看到，大多数的适用场所之间关联性非常强，说明大部分的香水产品不仅仅只有一个适用场所，而是有多个适用场所。例如，适合日常使用的香水，往往也适合在商务、party 聚会、约会上使用。在所有的 8 个适用场所中，日常、商务、party 聚会、约会出现次数最多，也是相互关联性最强的场所，另外 4 个场所(旅行、情趣、运动、其他)，则与其他适用场所关联性较小，说明这 4 个适用场所的香水产品针对性比较强。



图 5.21 过滤器节点变量设置



图 5.22 关联分析类型设置

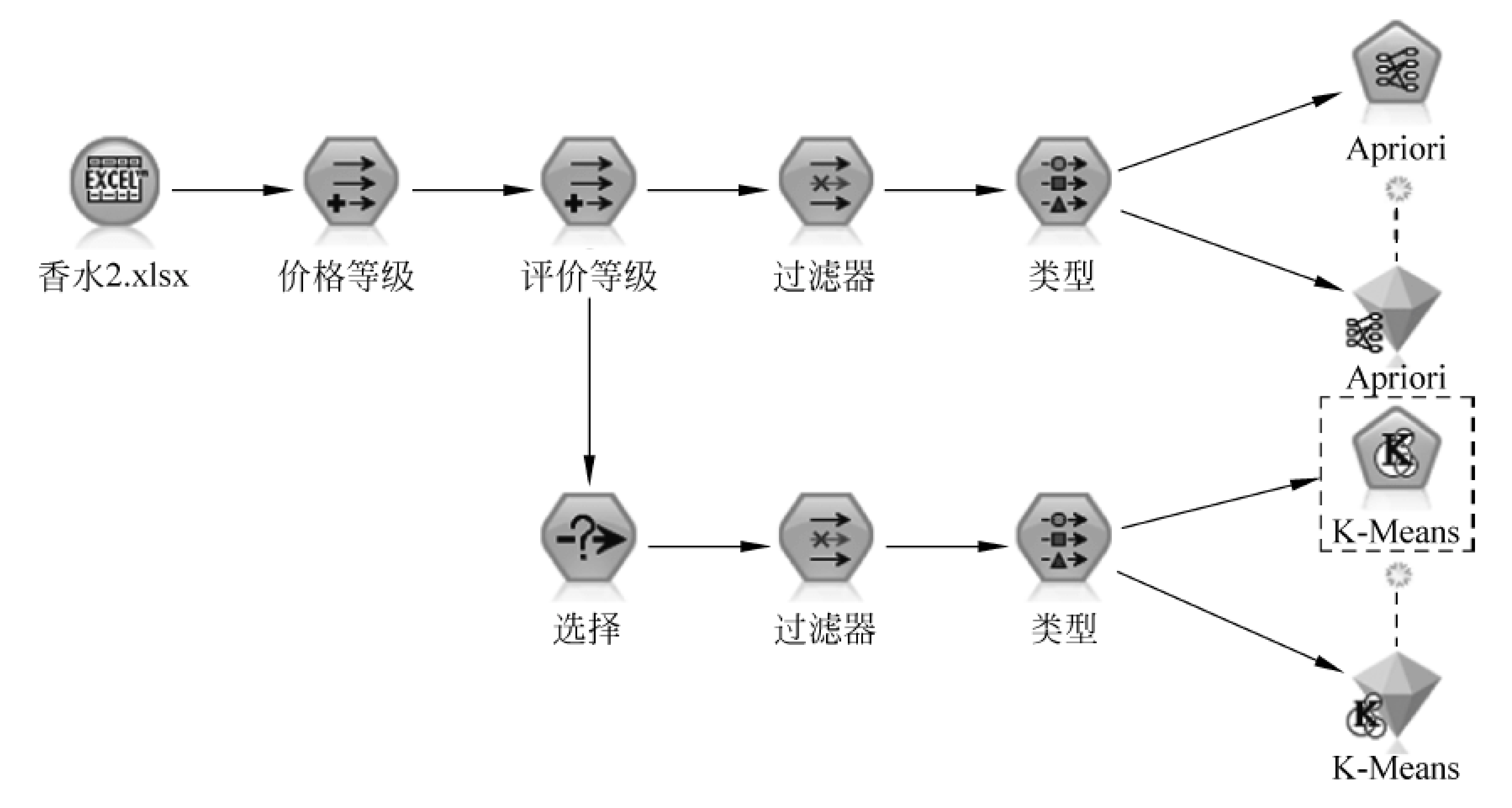


图 5.23 IBM SPSS Modeler 18.0 中具体的操作

排序依据: 置信度百分比 12 的 12

后项	前项	支持度百分比	置信度百分比
日常	商务	57.676	99.64
日常	party聚会 约会	56.432	98.897
日常	party聚会	60.373	98.625
约会	party聚会 日常	59.544	93.728
party聚会	商务	57.676	93.525
party聚会	商务 日常	57.469	93.502
约会	party聚会	60.373	93.471
日常	约会	66.805	93.168
约会	商务	57.676	91.727
约会	商务 日常	57.469	91.697
商务	party聚会 约会 日常	55.809	90.335
商务	party聚会 日常	59.544	90.244

图 5.24 适用场所关联分析结果

5.5 香水聚类分析

对香水进行聚类分析，将数据中的商品产地、包装、香调、净含量、分类、性别、适用场合数量作为输入字段，使用 SPSS Modeler 进行聚类分析。

图 5.25 适用过滤器节点过滤掉本次分析不需要的字段。图 5.26 将商品产地、包装、香调、净含量、分类、性别和适用场合数量作为输入进行聚类分析。

过滤器 注解

字段：已输入 20 个，已过滤 11 个，已重命名 0 个，已输出 9 个

字段	过滤器	字段
商品名称	×→	商品名称
商品产地	→	商品产地
包装	→	包装
香调	→	香调
净含量	→	净含量
分类	→	分类
性别	→	性别
价格	×→	价格
评价	×→	评价
旅行	×→	旅行
其他	×→	其他
约会	×→	约会
情趣	×→	情趣
商务	×→	商务
日常	×→	日常
party聚会	×→	party聚会
运动	×→	运动
适用场合数量	→	适用场合数量
价格等级	→	价格等级
评价等级	→	评价等级

图 5.25 过滤器节点变量设置

这里将聚类数确定为 6 的原因如下：如果聚类数设置为 5，那么最终得到的聚类质量较差，而且其中预测变量重要性最高的是香调，但得到的 5 个类别区分度不高，差异不明显。得到的 5 类香水包装和适用性别都是独立装和女，而且其中有一类的净含量值为空值，即该

类型 格式 注解

读取值 清除值 清除所有值

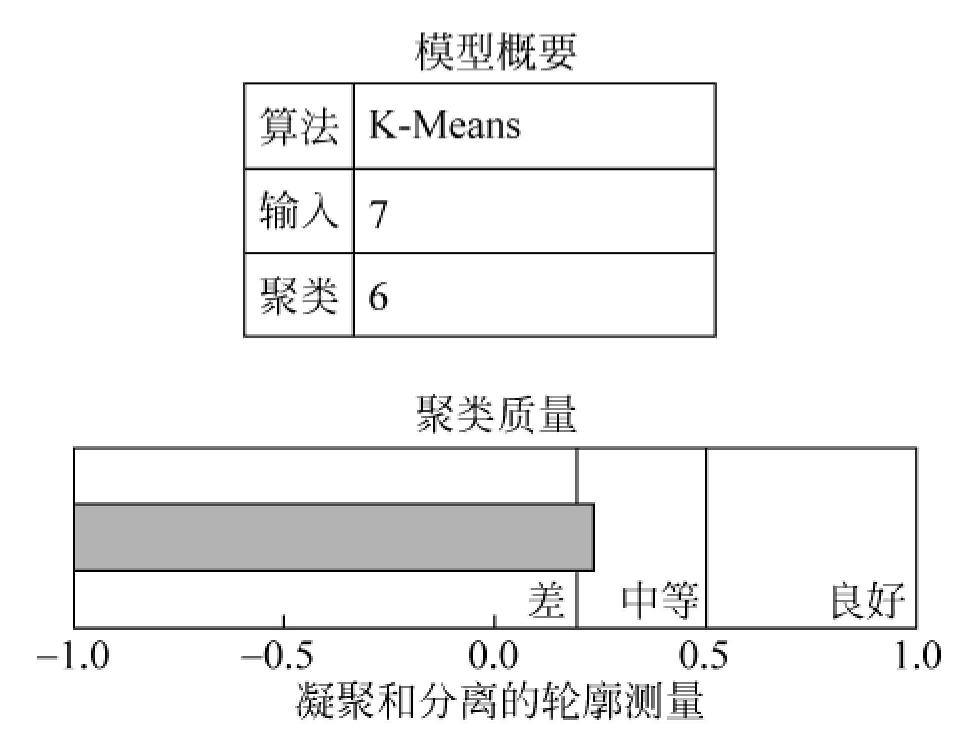
字段	测量	值	缺失	检查	角色
商品产地	名义	中国,德国,意大利,法国,美国,英...		无	输入
包装	名义	Q版香水,独立装,礼品套装,组合...		无	输入
香调	名义	东方香调,其他,木质香调,海洋香...		无	输入
净含量	名义	"101～200mL","16～30mL","1...		无	输入
分类	名义	其他,发香雾,古龙水,固体香水/香...		无	输入
性别	标记	通用/女		无	输入
适用场合数量	名义	0.0,1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0		无	输入
价格等级	名义	中等,低,较低,较高,非常高,高		无	无
评价等级	名义	中等,低,较低,较高,非常低,非常...		无	无

图 5.26　聚类类型节点设置

类的香水净含量分布均匀,并有明显的特征可供参考。

当聚类数设定为 7,甚至更多时,虽然聚类质量有所增加,但并不明显。最终得到的结果与聚类数为 6 得到的结果大致相同。而且如果聚类数过大,虽然聚类质量很好,但分类过细,会出现过拟合的情况,结果也没有意义。

进行聚类时,使用 K-Means 算法进行聚类,将聚类数设置为 6,即将数据中涉及的香水分为 6 个类别。聚类模型概要和聚类质量如图 5.27 所示。


模型概要

算法	K-Means
输入	7
聚类	6

图 5.27　聚类模型概要和聚类质量

如图 5.28 和图 5.29 所示,预测变量重要性依次为净含量、分类、包装、香调、商品产地、适用场合数量、性别。其中,净含量是聚类的主要依据,而性别则是聚类过程中,对结果影响最小的因素。

本次聚类的聚类质量为良好,平均 Silhouette 为 0.2。经过对数据的分析可知,在进行聚类时,数据分布不均。例如,同一种净含量规格的香水可能有多种香调,也可能来自不同产地,适用于不同场所,而聚类时不能兼顾净含量、香调、商品产地等多种因素,最终影响聚类结果。

经过对所有香水进行聚类分析,本次聚类分析中涉及的香水大致可以分为 6 类:

(1) 第一类:净含量为 31～100mL、淡香水 EDT、独立装、花果香调、产地为意大利、适用场合数量为 1、适用性别为女,所占比重为 24.5%。

(2) 第二类:净含量为 31～100mL、浓香水 EDP、独立装、花果香调、产地为法国、适用场合数量为 1、适用性别为女,所占比重为 23.0%。

(3) 第三类:净含量为 31～100mL、淡香水 EDT、独立装、花果香调、产地为法国、适用场合数量为 6、适用性别为女,所占比重为 22.1%。

聚类

输入(预测变量)重要性
■1.0 ■0.8 ■0.6 ■0.4 ■0.2 □0.0

聚类	聚类-1	聚类-4	聚类-3	聚类-6	聚类-2	聚类-5
标签						
描述						
大小	24.5% (112)	23.0% (105)	22.1% (101)	11.6% (53)	9.6% (44)	9.2% (42)
输入	净含量 31～100mL(66.1%)	净含量 31～100mL(98.1%)	净含量 31～100mL(74.3%)	净含量 31～100mL(79.2%)	净含量 1～15mL(90.9%)	净含量 16～30mL(97.6%)
	分类 淡香水EDT(100.0%)	分类 浓香水EDP(100.0%)	分类 淡香水EDT(100.0%)	分类 淡香水EDT(98.1%)	分类 淡香水EDT(79.5%)	分类 浓香水EDP(95.2%)
	包装 独立装(94.6%)	包装 独立装(96.2%)	包装 独立装(94.1%)	包装 独立装(92.5%)	包装 Q版香水(61.4%)	包装 独立装(92.9%)
	香调 花果香调(98.2%)	香调 花果香调(62.9%)	香调 花果香调(96.0%)	香调 混合香调(73.6%)	香调 花果香调(52.3%)	香调 花果香调(73.8%)
	商品产地 意大利(41.1%)	商品产地 法国(61.0%)	商品产地 法国(75.2%)	商品产地 法国(50.9%)	商品产地 法国(29.5%)	商品产地 法国(57.1%)
	适用场合数量 1.000(52.7%)	适用场合数量 1.000(22.9%)	适用场合数量 6.000(57.4%)	适用场合数量 1.000(30.2%)	适用场合数量 1.000(29.5%)	适用场合数量 1.000(33.3%)
	性别 女(96.4%)	性别 女(99.0%)	性别 女(94.1%)	性别 女(92.5%)	性别 女(90.9%)	性别 女(100.0%)

图 5.28　K-Means 聚类结果

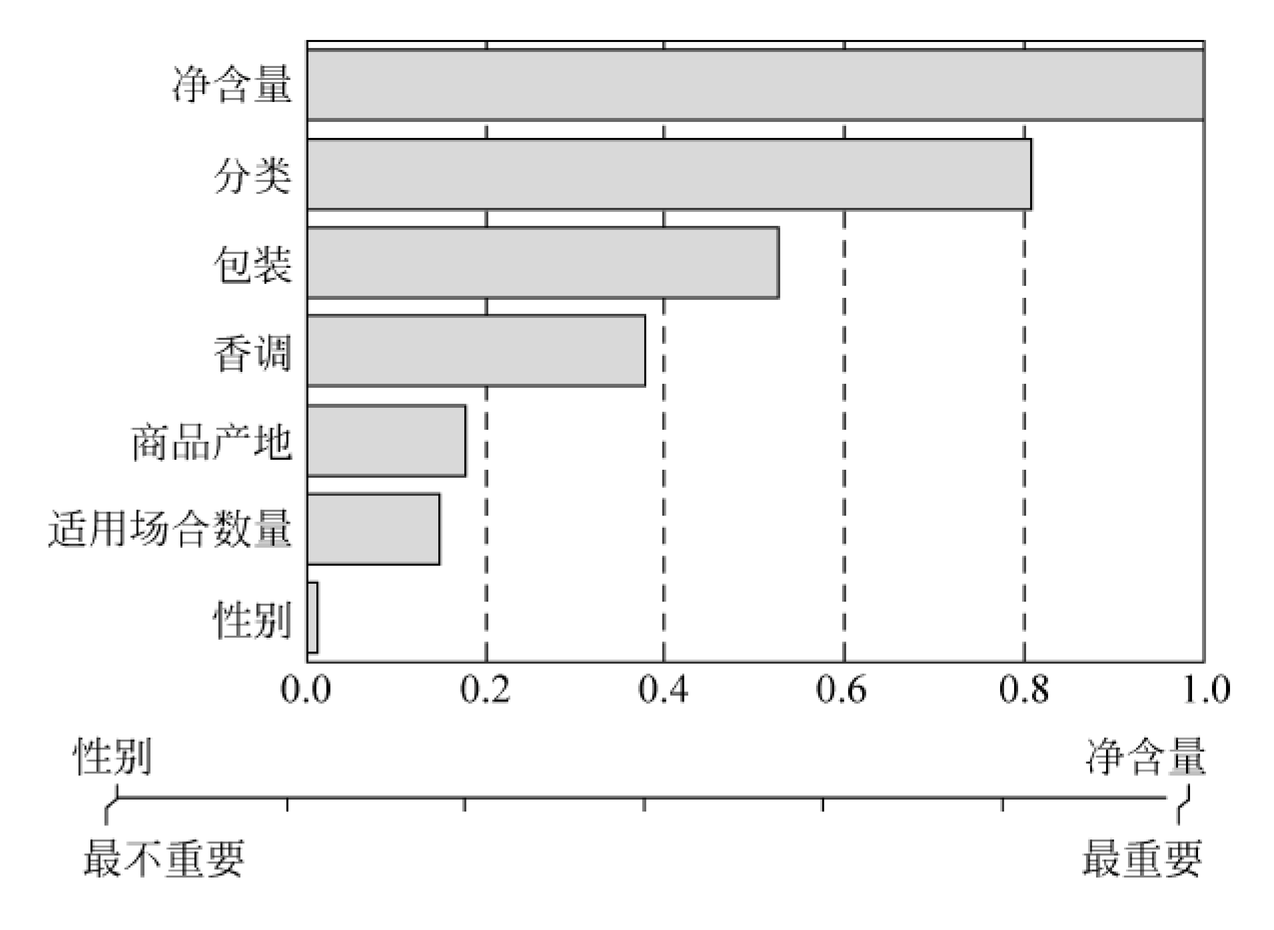


图 5.29　聚类预测变量重要性

(4) 第四类：净含量为 31～100mL、淡香水 EDT、独立装、混合香调、产地为法国、适用场合数量为 1、适用性别为女，所占比重为 11.6%。

(5) 第五类：净含量为 1～15mL、淡香水 EDT、Q 版香水、花果香调、产地为法国、适用场合数量为 1、适用性别为女，所占比重为 9.6%。

(6) 第六类：净含量为 16～30mL、浓香水 EDP、独立装、花果香调、产地为法国、适用场合数量为 1、适用性别为女，所占比重为 9.2%。

从净含量的角度来看，大部分香水都是 31～100mL，所占比重超过 80%。从分类角度来看，大部分香水为淡香水 EDT；从包装角度来看，大部分香水是独立装，只有净含量为 1～15mL 的香水是 Q 版香水；从香调的角度来看，大部分香水是花果香调，混合香调的香水在总类别中占比重较小；从商品产地的角度来看，大部分香水产自法国，产自其他国家的香水较少；从适用场合数量来看，大部分香水适用场合数量为 1，说明大部分香水都适用于固定的场合，而有些香水适用场合数量为 6，但所占比重不大，说明普遍适用的香水种类不是很多。从性别角度来看，几乎所有香水都适用于女性，男女通用的香水种类较少，也说明进行聚类时，性别对聚类结果的影响较低。

5.6 香水营销建议

结合上述分析，对于希望提升销量的商家来说，在销售的香水产品选择上，需要选择消费者欢迎、总体销量好的产品。

(1) 制定价格方面。商家可以将产品价格定位在大众消费品的水平上，并保持正常利润空间，更多考虑运用价格策略扩大产品销路，吸引更多的消费者。结合香水产品的分类来看，淡香水 EDT 的销量与价格呈负相关；而浓香水 EDP 的销量与价格呈正相关。说明浓香水 EDP 的买家比较注重品质，对价格不敏感，而淡香水 EDT 的买家对价格敏感。制定价格时，对于淡香水 EDT 类的产品，可以根据消费者的心理价格，在不亏损的前提下，适当降低产品的价格，从而获取更多客户，达到利润最大化；对于浓香水 EDP 类产品，消费者追求奢侈品牌，价格越高越能刺激其需求，在调整空间内，提高浓香水 EDP 产品的价格，刺激这类追求品质的消费者购买，可以获取更多客户，从而增加收入。

(2) 产品分类方面。香水产品的产地、香调、净含量都会对销售产生很大的影响，因此选择正确类别的香水产品进行销售是提升销量非常重要的方面。法国、意大利是世界上重要的香水奢侈品产地，法国、意大利的香水产品在世界范围内有着巨大的影响力；国产香水在价格上有着更大的优势，口味也更符合我国消费者的喜好。在香调方面，我国消费者喜好清淡的口味，因此花果香调之类的清新口味有更大的市场。在净含量方面，便携性好的小包装香水产品更受消费者青睐。商家选择销售的产品时，需要综合考虑产地、香调、净含量，选择更受消费者欢迎的产品才能获得更多收入。

(3) 销售策略方面。由于消费者在购买香水产品时体现出了明显的价格敏感性，价格低的香水产品销量更好。组合装的香水销量好于其他包装。另外，目前我国香水消费者中很大一部分还是购买香水作为礼品。因此，商家为了吸引更多消费者，可以制定一个短期促销策略，降低香水产品的价格，通过价格优势吸引消费者的注意力，并且推出更多的香水组合以及礼品装香水，结合不同适用场合的消费需求，满足不同消费者群体，刺激特定消费者群体消费。

第 6 章

银行信用卡欺诈与拖欠行为分析

信用卡作为一种全新的支付手段和信用工具，是中国个人金融服务市场中成长最快的产品线之一。信用卡能够给银行带来很高的利润。目前，我国信用卡透支贷款的年利率为18%左右，同时还会带来相当可观的分期付款手续费收入和商户回佣等中间业务收入。根据各上市银行 2015 年年报、公开媒体的数据统计，截至 2015 年年末，全国信用卡累计发卡 4.32 亿张，授信总额为 7.08 万亿元，截至 2016 年年末，工商银行发卡量超过 1.2 亿张，建设银行和招商银行发卡量分别为 9407 万张和 8031 万张。根据中国银联官方数据，2015 年银行信用卡业务总收入为 649.03 亿元，同比增长 38.11%。信用卡业务给银行带来高收益的同时，也伴随着高风险。截至 2015 年末，我国信用卡业务逾期半年未偿信贷总额为 380.27 亿元，较 2014 年增加 22.63 亿元，增长率为 6.33%。我国的信用卡平均不良率已经达到 2.07%，为近年来新高。

我国的信用卡业务较国外起步晚，与国外成熟的信用卡市场相比其规模还很小，相关的制度还不够完善。作为纯信用模式下的金融信贷产品，信用卡风险主要包括三个方面：信用风险、欺诈风险、操作风险。近年来，随着互联网金融的快速发展以及支付模式日益多元化，信用卡违约现象逐渐增多，不良贷款快速增长，信用卡欺诈、违法套现等违法犯罪活动不断出现，并呈现出新趋势、新特点。信用卡欺诈不仅给银行造成经济损失，还会带来巨大的声誉风险，降低客户对银行的信任度。对此，各银行加强信用卡管理，提升风险防控能力已经刻不容缓。

本案例获取某银行的客户信用卡记录，挖掘数据的潜在价值，为该银行的信用卡业务决策提供参考。该银行面临的信用卡欺诈和拖欠现象比较严重，发生比例高于我国银行行业的平均值。本案例希望通过对影响用户信用等级的主要因素进行分析，以及结合信用卡用户的人口特征属性对欺诈行为和拖欠行为的影响因素进行分析。

通过对银行的客户信用记录、申请客户信息、拖欠历史记录、消费历史记录等数据进行分析，对不同信用程度的客户进行归类，研究信用卡贷款拖欠、信用卡欺诈等问题与客户的

个人信息、信用卡使用信息的关系，为银行提前识别、防控信用卡业务风险提供参考，从而减少银行在信用卡业务方面的损失。

6.1 客户信用等级影响因素

个人信用卡的信用风险是指借款人不能在规定期限内按照约定及时、足额偿还银行本金和利息。随着信用卡使用的日益广泛，申请信用卡的客户增多，也给银行带来了更大的潜在信用风险，银行需要采取相应措施，规避或是减轻个人信用卡的信用风险。

对申请新信用卡的个人用户进行信用分析和等级评定，是银行控制个人信用卡信用风险的一项必要措施。在客户向银行申请信用卡时，银行会根据用户提供的个人信息进行评分，综合考虑客户的各项指标，对每一项指标都按照一定的标准评分，然后累计得到客户的信用总评分，为每位客户制定信用等级，给予相应的信用卡额度。对潜在价值高且信用风险低的客户，给予大的信用额度；而对潜在价值低或信用风险高的用户，给予小的额度。

6.1.1 客户信用卡申请数据预处理

在客户申请信用卡时，主要考虑因素见表 6.1。

表 6.1 用户信用等级评价指标

一级指标	个人自然情况	个人职业情况	个人收入及财产	个人银行记录
二级指标	年龄	职业类别	年收入	信贷情况
	性别	工作年限	居住类型	
	户籍		车辆情况	
	婚姻状态		保险缴纳	
	教育程度			

从银行获取的个人信用卡客户相关数据中选取“申请客户信息”和“客户信用记录”两个表格，在 SPSS Modeler 18.0 中按照关键词“客户号”进行合并，删除重复字段。由于“申请客户信息”中未申请成功的用户在“客户信用记录”中没有相应的信用等级相关记录，信用总评分、信用等级、额度、审批结果显示为 null，如图 6.1 所示。

客户号	信用总评分	信用等级	额度	审批结果	...	...	年龄	性别	婚姻状态	教育程度	职业类别
000099994...	$null$	$null$	$null$	$null$	...	...	45....	女	未婚	本科	私营企业
000099994...	$null$	$null$	$null$	$null$	...	...	21....	男	未婚	大专	外资企业
000099994...	$null$	$null$	$null$	$null$	...	...	22....	女	未婚	大专	私营企业
000099996...	86.000	B-良好客户	50000.0...	0-通过	...	...	25....	男	未婚	本科	个体户
000099997...	90.000	A-优质客户	100000....	0-通过	...	...	25....	男	未婚	本科	私营企业
000099998...	86.000	B-良好客户	50000.0...	0-通过	...	...	25....	女	未婚	本科	私营企业
000099998...	84.000	B-良好客户	50000.0...	0-通过	...	...	32....	男	未婚	本科	私营企业
000099998...	87.000	B-良好客户	50000.0...	0-通过	...	...	52....	女	未婚	本科	私营企业
000099998...	$null$	$null$	$null$	$null$	...	...	32....	男	未婚	大专	私营企业
000099998...	$null$	$null$	$null$	$null$	...	...	41....	女	已婚	本科	个体户

图 6.1 合并完成后的用户信息记录

对没有值的字段进行填充，将合并完成后的表格完善，便于后面对影响用户信用等级的因素进行分析。通过使用“填充”节点，具体处理方法如图 6.2 所示，统一将没通过审批的用

户信用总评分设置为 0，信用等级为“F-未通过客户”，额度为 0，审批结果为“1-未通过”。



图 6.2 将信用等级为空值的字段填充替换为未通过客户

数据分析过程中并不需要客户的个人标识信息，使用“过滤器”节点，将“客户号”“客户姓名”“证件号码”等标识用户的变量过滤，由于“额度”“信用总评分”变量和“信用等级”变量作用重复，并且对应关系明确，因此将其删除，如图 6.3 所示。

过滤器 注解

字段：已输入 19 个，已过滤 6 个，已重命名 0 个，已输出 13 个

字段	过滤器	字段
客户号	✕→	客户号
信用总评分	✕→	信用总评分
信用等级	→	信用等级
额度	✕→	额度
审批结果	✕→	审批结果
客户姓名	✕→	客户姓名
证件号码	✕→	证件号码
年龄	→	年龄
性别	→	性别
婚姻状态	→	婚姻状态

◉ 查看当前字段 ○ 查看未使用的字段设置

图 6.3 “过滤器”节点属性设置

使用“类型”节点，将“信用等级”字段设置为目标，其他与客户有关的个人信息字段设置为输入，使用 C5.0 决策树算法分析用户的人口属性对用户信用等级的影响，设置页面如图 6.4 所示。



字段	测量	值	缺失	检查	角色
信用等级	名义	A-优质客...		无	目标
年龄	连续	[18.0,80.0]		无	输入
性别	标记	男/女		无	输入
婚姻状态	名义	丧偶,已婚,...		无	输入
教育程度	名义	初中及以...		无	输入
职业类别	名义	个体户,其...		无	输入
户籍	名义	上海,内蒙...		无	输入
居住类型	名义	其他,租房,...		无	输入

图 6.4 “类型”节点属性设置

对所有审批通过的用户，信用评分为 60～69 分的用户对应的信用等级为“D-风险客户”，相应的信用卡为“普卡”，额度为 10 000 元；信用评分为 70～79 分的客户对应的信用等级为“C-普通客户”，相应的信用卡为“银卡”，额度为 20 000 元；信用评分为 80～89 分的用户对应的信用等级为“B-良好客户”，相应的信用卡为“金卡”，额度为 50 000 元；信用评分为 90～100 分的用户对应的信用等级为“A-优质客户”，相应的信用卡为“白金卡”，额度为 100 000 元。

6.1.2 信用卡申请成功影响因素

在信用卡申请的审批过程中，需要区分某些潜在价值低且信用风险高的客户，拒绝某些指标达不到要求的申请，为了方便信用卡中心对申请记录进行量化审批，对所有申请记录和最终获批的客户列表进行关联分析，得到信用卡能否申请成功的主要影响因素，供信用卡中心参考。

图 6.5 是申请信用卡能否成功的影响因素分析流程，分别使用线性支持向量机 SVM 和 SVM 模型进行分析，并使用逻辑回归计算各变量的相关系数。使用分区节点将所有数据按照训练集 70％和测试集 30％的比例分配记录。

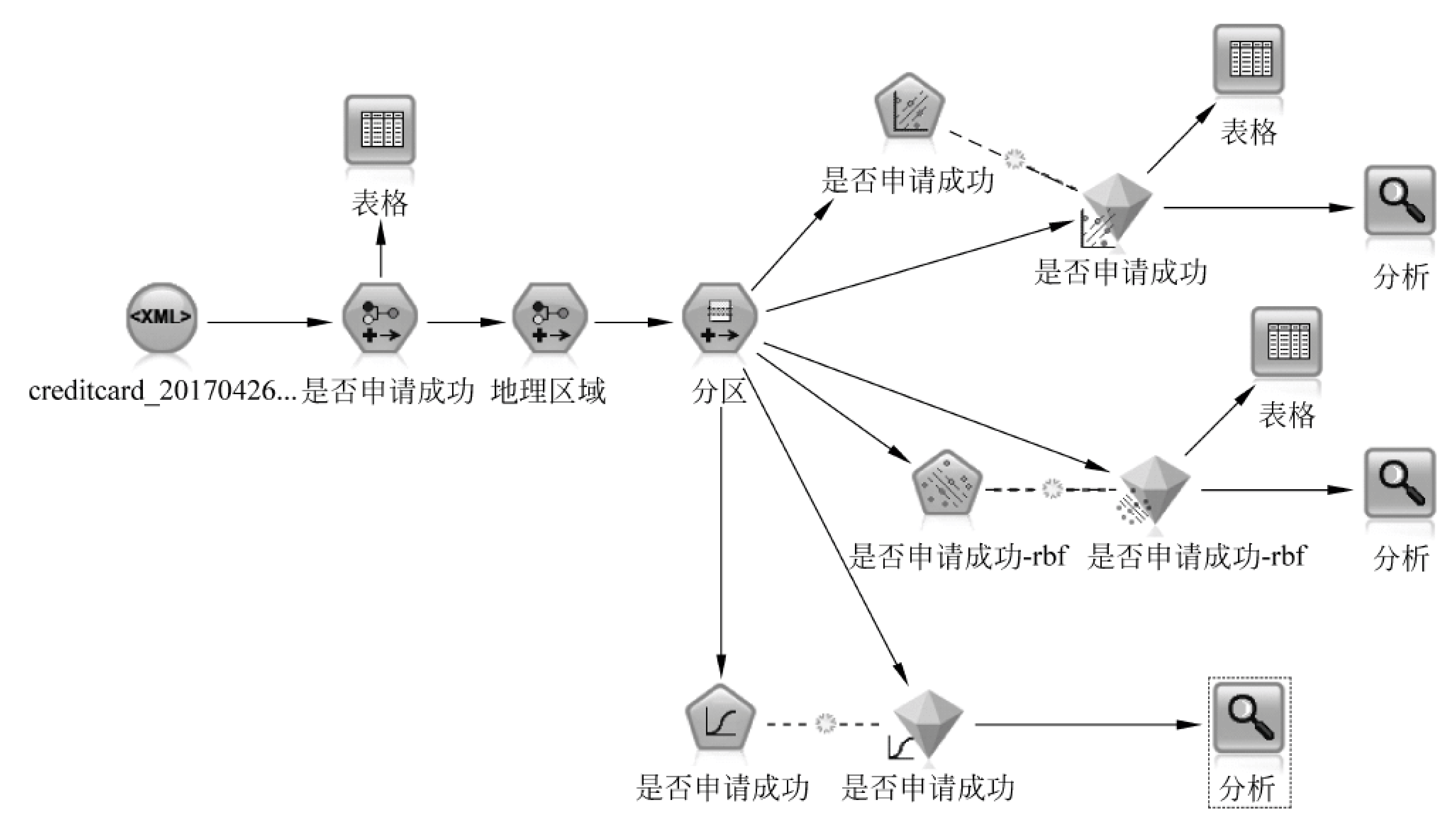


图 6.5 申请信用卡成功与否影响因素分析

数据预处理之后以信用等级中“F-未通过客户”表示未通过的用户，将其设置为申请失败用户，将所有 A～D 信用等级的用户统一置为申请成功，如图 6.6 所示。

应用线性 SVM 模型对年收入、信贷情况、保险缴纳、车辆情况、教育程度等进行分析，并计算各变量的预测变量重要性，在线性 SVM 的模型结果后放置表格节点，显示模型的结果值，如图 6.7 所示，可以看到“＄LC-是否申请成功”列中显示了预测成功的概率。

模型评价分析结果如图 6.8 所示，训练过程中准确率为 88.58％，应用测试集进行验证，线性 SVM 达到 89.68％的分类准确性。在申请成功的记录中，分类正确的记录数达到 3723 条，占总数的 89％，失败的条数为 475 条，占总数的 11％。



图 6.6 申请结果重新分类

	年龄	性别	婚姻	教育程度	职业	户籍	居住类型	车辆情况	保险缴纳	工作年限	年收入	信贷情况	信用等级	是否申请成功	分区	$L-是否申请成功	$LC-是否申请成功
1	60	女	已婚	本科	私营企业	北京	自购房	有	有	37	22297	现在没有贷款	D-风险客户	成功	1_培训	成功	0.608
2	32	男	已婚	硕士及以上	私营企业	湖南	自购房	有	有	7	19360	正常还款	D-风险客户	成功	1_培训	成功	0.671
3	45	女	已婚	本科	私营企业	上海	自购房	有	有	22	22599	正在偿还	D-风险客户	成功	2_测试	成功	0.594
4	29	男	未婚	硕士及以上	私营企业	天津	自购房	有	有	4	22975	现在没有贷款	D-风险客户	成功	2_测试	成功	0.632
5	29	女	未婚	硕士及以上	私营企业	宁夏	自购房	有	有	4	22975	现在没有贷款	D-风险客户	成功	1_培训	成功	0.619
6	46	女	已婚	本科	私营企业	重庆	自购房	有	有	23	23000	现在没有贷款	D-风险客户	成功	1_培训	成功	0.571
7	46	男	未婚	本科	私营企业	四川	自购房	有	有	23	19360	正在偿还	D-风险客户	成功	2_测试	成功	0.620
8	59	男	已婚	本科	私营企业	天津	自购房	有	有	36	23000	正在偿还	D-风险客户	成功	1_培训	成功	0.646
9	51	女	已婚	本科	私营企业	宁夏	自购房	有	有	28	20322	现在没有贷款	D-风险客户	成功	1_培训	成功	0.598
10	31	女	已婚	硕士及以上	私营企业	湖北	自购房	有	有	6	23420	没有贷款记录	D-风险客户	成功	1_培训	成功	0.566
11	68	女	已婚	本科	私营企业	广西	自购房	有	有	45	23460	现在没有贷款	D-风险客户	成功	2_测试	成功	0.661
12	48	男	未婚	大专	私营企业	陕西	自购房	有	有	28	23476	现在没有贷款	D-风险客户	成功	2_测试	成功	0.683
13	45	女	已婚	本科	私营企业	天津	自购房	有	有	22	21590	正在偿还	D-风险客户	成功	2_测试	成功	0.578
14	54	女	未婚	本科	个体户	湖北	自购房	有	有	31	21681	正在偿还	D-风险客户	成功	1_培训	成功	0.615
15	46	男	已婚	大专	外资企业	四川	自购房	有	有	26	21825	现在没有贷款	D-风险客户	成功	2_测试	成功	0.681
16	47	女	已婚	大专	私营企业	广西	自购房	有	有	27	22058	正常还款	D-风险客户	成功	2_测试	成功	0.670
17	64	男	未婚	大专	外资企业	天津	自购房	有	有	44	23765	正常还款	D-风险客户	成功	1_培训	成功	0.724
18	31	女	未婚	硕士及以上	私营企业	甘肃	自购房	有	有	6	23916	现在没有贷款	D-风险客户	成功	1_培训	成功	0.604
19	45	男	离异	本科	私营企业	浙江	自购房	有	有	22	23980	现在没有贷款	D-风险客户	成功	2_测试	成功	0.600
20	51	女	已婚	本科	外资企业	重庆	自购房	有	有	28	24108	正在偿还	D-风险客户	成功	1_培训	成功	0.593
21	56	女	离异	本科	个体户	福建	自购房	有	有	33	24113	现在没有贷款	D-风险客户	成功	1_培训	成功	0.617
22	56	女	未婚	大专	外资企业	四川	自购房	有	有	36	24190	现在没有贷款	D-风险客户	成功	2_测试	成功	0.694
23	66	男	未婚	本科	国有企业	湖南	自购房	有	有	43	25500	正在偿还	D-风险客户	成功	1_培训	成功	0.689
24	48	男	已婚	本科	私营企业	北京	自购房	有	有	25	10634	现在没有贷款	D-风险客户	成功	2_测试	成功	0.577
25	29	男	未婚	硕士及以上	外资企业	青海	自购房	有	有	4	25575	现在没有贷款	D-风险客户	成功	1_培训	成功	0.611
26	42	男	未婚	大专	国有企业	黑...	自购房	有	有	22	10810	现在没有贷款	D-风险客户	成功	1_培训	成功	0.625
27	59	男	已婚	大专	私营企业	山东	自购房	有	有	39	10860	正在偿还	D-风险客户	成功	2_测试	成功	0.691
28	52	男	已婚	大专	私营企业	广东	自购房	有	有	32	11230	现在没有贷款	D-风险客户	成功	2_测试	成功	0.692
29	47	女	已婚	本科	私营企业	贵州	自购房	有	有	24	11320	现在没有贷款	D-风险客户	成功	1_培训	成功	0.558

图 6.7 线性 SVM 分析结果

线性 SVM 模型中各变量的重要性如图 6.9 所示，其中年收入的重要性最高，重要性超过了 0.7，其次是信贷情况、保险缴纳、车辆情况、教育程度、户籍、工作年限、职业、年龄等变量较不重要，婚姻变量作用不显著。

为了对比，使用 SVM 模型进行分析，模型使用专家模式，应用 RBF 内核类型，计算 SVM 预测变量的重要性，如图 6.10 所示。

结果与线性 SVM 模型具有较大差异，特别是年收入这一项，在线性 SVM 模型中排名最靠前，但在 SVM 模型中排名靠后，应用分析节点对 SVM 模型的结果进行分析，如

模型信息

目标字段	是否申请成功
模型构建方法	线性 SVM
输入的预测变量数	12
最终模型中的预测变量数	11
规则化类型	L2
惩罚参数 (Lambda)	0.100
分类准确性	88.6%

混淆矩阵

实测	预测 成功	失败	比例正确
成功	3,723	475	0.89
失败	327	2,499	0.88
比例正确	0.92	0.84	0.89

输出字段 是否申请成功 的结果

比较 $L-是否申请成功 与 是否申请成功

"分区"	1_培训		2_测试	
正确	6,222	88.58%	2,669	89.68%
错误	802	11.42%	307	10.32%
总计	7,024		2,976	

图 6.8 线性 SVM 模型综合结果

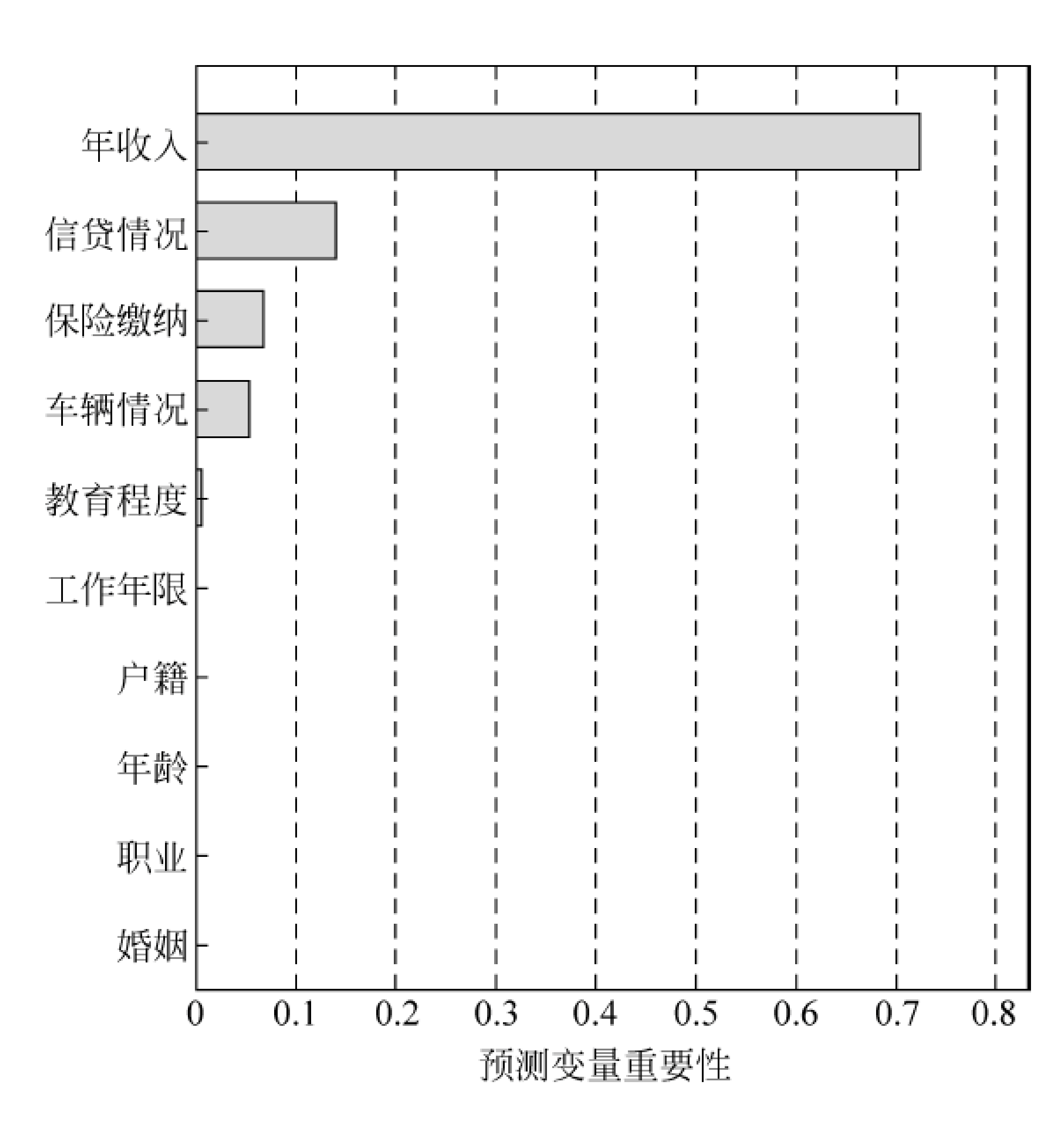


图 6.9 线性 SVM 模型中各变量的重要性

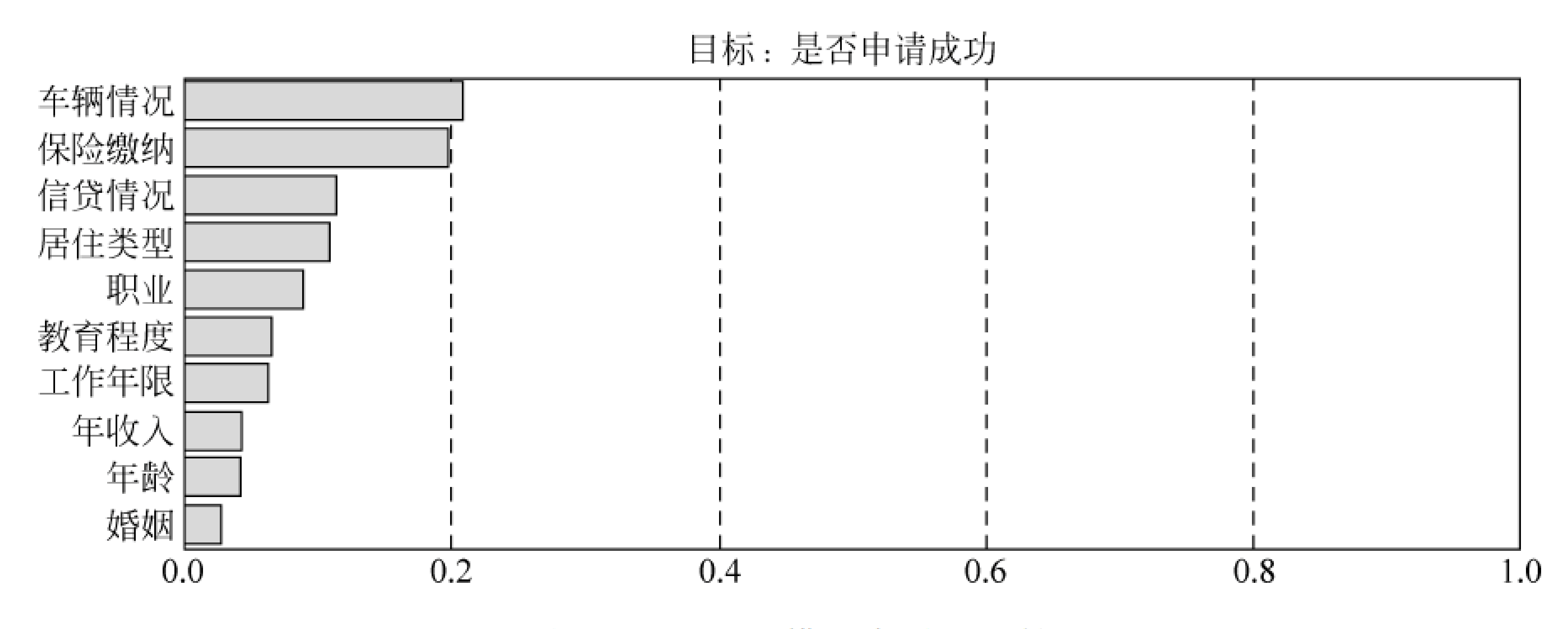


图 6.10 SVM 模型变量重要性

图 6.11 所示，其测试集的准确率只有 65.89%，低于线性 SVM 模型的 89.68%。

综上，在实际应用中建议使用线性 SVM 模型进行用户信用的影响因素分析，在用户申请信用卡过程中依次使用年收入、信贷情况、保险缴纳、车辆情况、教育程度、户籍、工作年限、职业、年龄、性别等进行评价。

输出字段 是否申请成功 的结果

比较 $S-是否申请成功 与 是否申请成功

'分区'	1_培训		2_测试	
正确	5,417	77.12%	1,961	65.89%
错误	1,607	22.88%	1,015	34.11%
总计	7,024		2,976	

图 6.11 SVM 模型结果分析

为了将各项变量指标进行定量分析，使用逻辑回归对各影响因素进行分析，并对户籍进行向上钻取，按照地理区域进行划分，如“华东”包括山东、江苏、上海、浙江、安徽、江西 6 个省市，运行模型后得到的结果如图 6.12 所示，重要性指标与线性 SVM 模型大致相同。

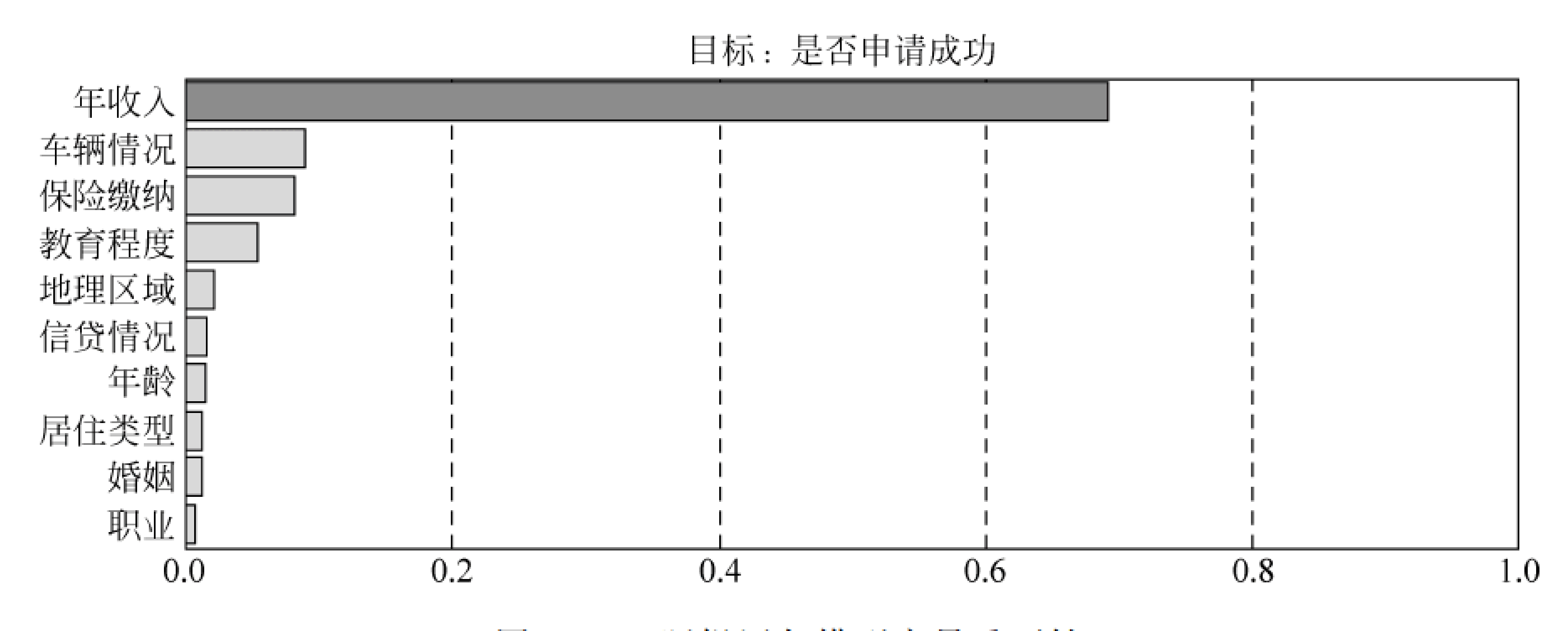


图 6.12 逻辑回归模型变量重要性

其中，表 6.2 是各项影响因素变量的分布情况，包括各个分类输入变量的数量及所占总记录数的比例。

表 6.2 各影响因素变量的分布情况

		N	Marginal Percentage
是否申请成功	成功	4198	59.8%
	失败	2826	40.2%
婚姻	离异	324	4.6%
	丧偶	12	0.2%
	未婚	4573	65.1%
	已婚	2115	30.1%
教育程度	本科	3403	48.4%
	初中及以下	709	10.1%
	大专	1608	22.9%
	高中	693	9.9%
	硕士及以上	611	8.7%
居住类型	其他	92	1.3%
	自购房	1492	21.2%
	租房	5440	77.4%
车辆情况	无	5296	75.4%
	有	1728	24.6%

续表

		N	Marginal Percentage
保险缴纳	无	3197	45.5%
	有	3827	54.5%
信贷情况	还在拖欠	186	2.6%
	没有贷款记录	145	2.1%
	现在没有贷款	3741	53.3%
	逾期还款	30	0.4%
	正常还款	794	11.3%
	正在偿还	2128	30.3%
职业	个体户	920	13.1%
	国有企业	479	6.8%
	其他企业	477	6.8%
	私营企业	4192	59.7%
	外资企业	956	13.6%
性别	男	4926	70.1%
	女	2098	29.9%
地理区域	东北	690	9.8%
	华北	1335	19.0%
	华东	1826	26.0%
	华南	800	11.4%
	华中	653	9.3%
	西北	971	13.8%
	西南	749	10.7%
Valid		7024	100.0%
Missing		0	
Total		7024	
Subpopulation		7011	

说明：上述表格是由 SPSS modeler 自动生成的，其中 N 表示数量，Marginal Percentage 表示所占比例。

模型结果的拟合情况如图 6.13 所示，其 Sig 指标为 0 说明模型具有较高的显著性。

Model Fitting Information

	Model Fitting Criteria			Likelihood Ratio Tests		
Model	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	9469.608	9476.465	9467.608			
Final	1874.781	2087.351	1812.781	7654.827	30	.000

图 6.13　逻辑回归模型拟合情况

模型的因变量虚拟回归系数如图 6.14 所示，其中 Cox and Snell 指标为 0.664，Nagelkerke 参数为 0.897，McFadden 参数为 0.809，说明逻辑回归模型的质量较好。

Pseudo R-Square

Cox and Snell	.664
Nagelkerke	.897
McFadden	.809

图 6.14　逻辑回归模型变异情况

使用分析节点对结果进行分析，其中训练集准确率达到94.01%，测试集的准确率为95.16%，如图6.15所示，说明逻辑回归具有较高的应用价值。

输出字段 是否申请成功 的结果

比较 $L-是否申请成功 与 是否申请成功

'分区'	1_培训		2_测试	
正确	6,603	94.01%	2,832	95.16%
错误	421	5.99%	144	4.84%
总计	7,024		2,976	

$LP-是否申请成功 的置信度值报告

'分区'= 1_培训	
范围	0.5 - 1.0
平均正确性	0.959
平均不正确性	0.713
正确性始终高于	0.977 （观测值的 70.67%）
不正确性始终低于	0.501 （观测值的 0.03%）
94.01% 以上的准确性	0.0
2.0 以上的折叠正确性	0.732 （观测值的 97.01%）
'分区'= 2_测试	
范围	0.5 - 1.0
平均正确性	0.958
平均不正确性	0.721
正确性始终高于	0.979 （观测值的 70.13%）
不正确性始终低于	0.5 （观测值的 0%）
95.16% 以上的准确性	0.0
2.0 以上的折叠正确性	0.731 （观测值的 97.6%）

图6.15 逻辑回归模型结果对比分析

将逻辑回归结果以回归方程的形式进行量化，结果如图6.16所示，用户申请信用卡时将其提交的资料应用于回归方程中，可得到审批结果。

方程式用于 成功

0.06546 * 工作年限 +
0.0001316 * 年收入 +
0.03073 * 年龄 +
-0.3525 * [婚姻=离异] +
-0.152 * [婚姻=丧偶] +
-0.1204 * [婚姻=未婚] +
-2.714 * [教育程度=本科] +
-3.755 * [教育程度=初中及以下] +
-2.622 * [教育程度=大专] +
-3.14 * [教育程度=高中] +
-0.6427 * [居住类型=其他] +
0.3434 * [居住类型=自购房] +
-2.988 * [车辆情况=无] +
-4.497 * [保险缴纳=无] +
-1429.4 * [信贷情况=还在拖欠] +
-0.434 * [信贷情况=没有贷款记录] +
-0.2581 * [信贷情况=现在没有贷款] +
-1213.6 * [信贷情况=逾期还款] +
-0.0532 * [信贷情况=正常还款] +
-0.02383 * [职业=个体户] +
-0.1395 * [职业=国有企业] +
-0.2731 * [职业=其他企业] +
-0.2375 * [职业=私营企业] +
-0.4284 * [性别=男] +
-0.1751 * [地理区域=东北] +
-0.02093 * [地理区域=华北] +
-0.3923 * [地理区域=华东] +
0.07523 * [地理区域=华南] +
-0.1488 * [地理区域=华中] +
-0.409 * [地理区域=西北] +
+ -2.081

图6.16 逻辑回归方程结果

6.2 信用卡客户信用等级影响因素

在 SPSS Modeler 18.0 中的处理,具体流程图如图 6.17 所示。

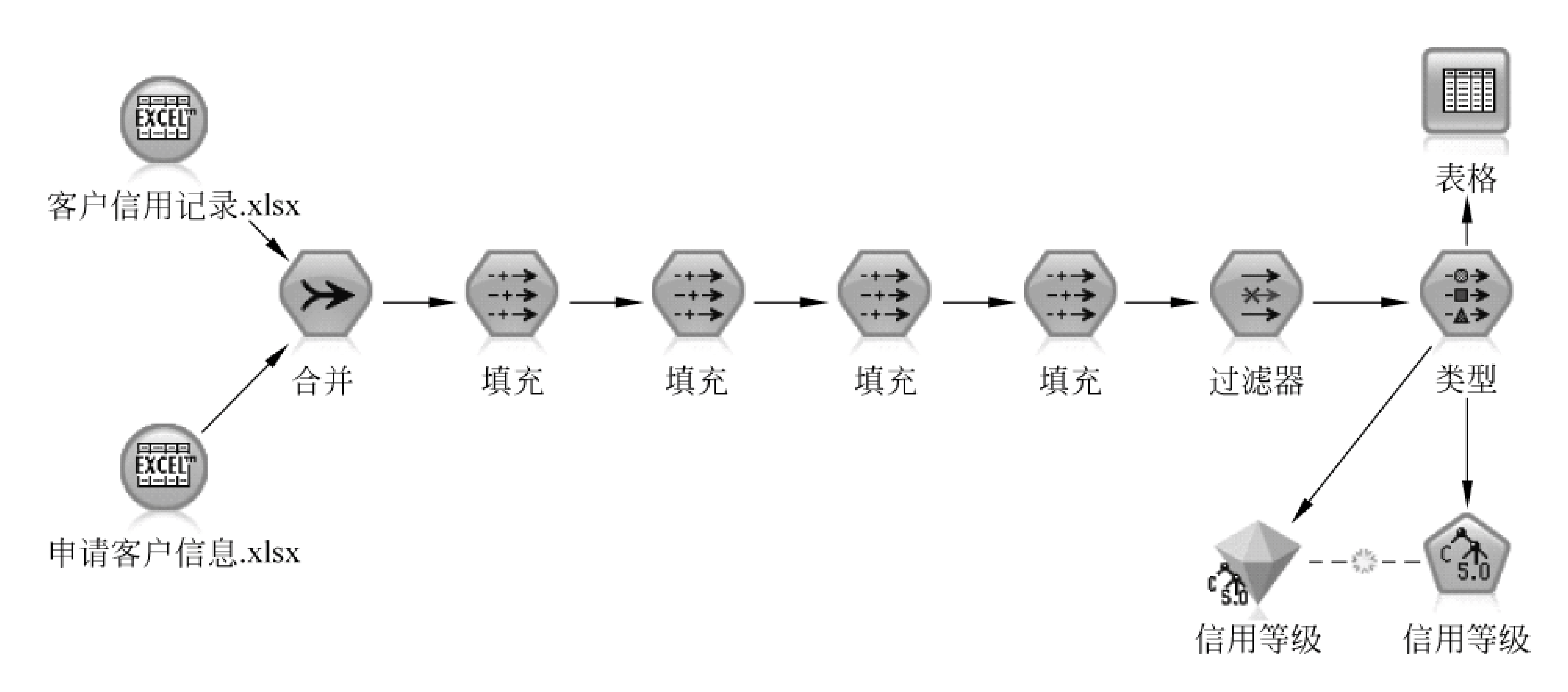


图 6.17 信用等级影响因素分析流程图

C5.0 决策树算法是应用于较大数据集上的分类算法,在执行效率和内存使用方面进行了改进;在面对数据遗漏和输入字段很多的问题上非常稳健,而且通常不需要很多的训练次数进行估计,在训练的时候提高了运行效率;相比其他类型的模型,更容易理解,模型推出的规则有非常直观的解释。采用 C5.0 决策树算法对决定用户信用等级的因素进行分析,挖掘银行对个人用户信用等级进行评价时的影响因素及相应的重要性。

在 SPSS Modeler 18.0 中以信用等级为目标,其他所有变量为输入,运行 C5.0 决策树算法,得到模型的变量重要性分布情况如图 6.18 所示。在预测变量重要性分布图中,可以看到银行在评判个人用户的信用等级时,最重要的评价因素是用户的年收入,重要性远超过其他变量,次重要的因素是用户的居住类型,其次是教育程度、车辆情况、年龄、保险缴纳、单笔消费金额,日均消费次数、工作年限等,与年收入和居住类型相比,其他变量之间的重要性差异较小。

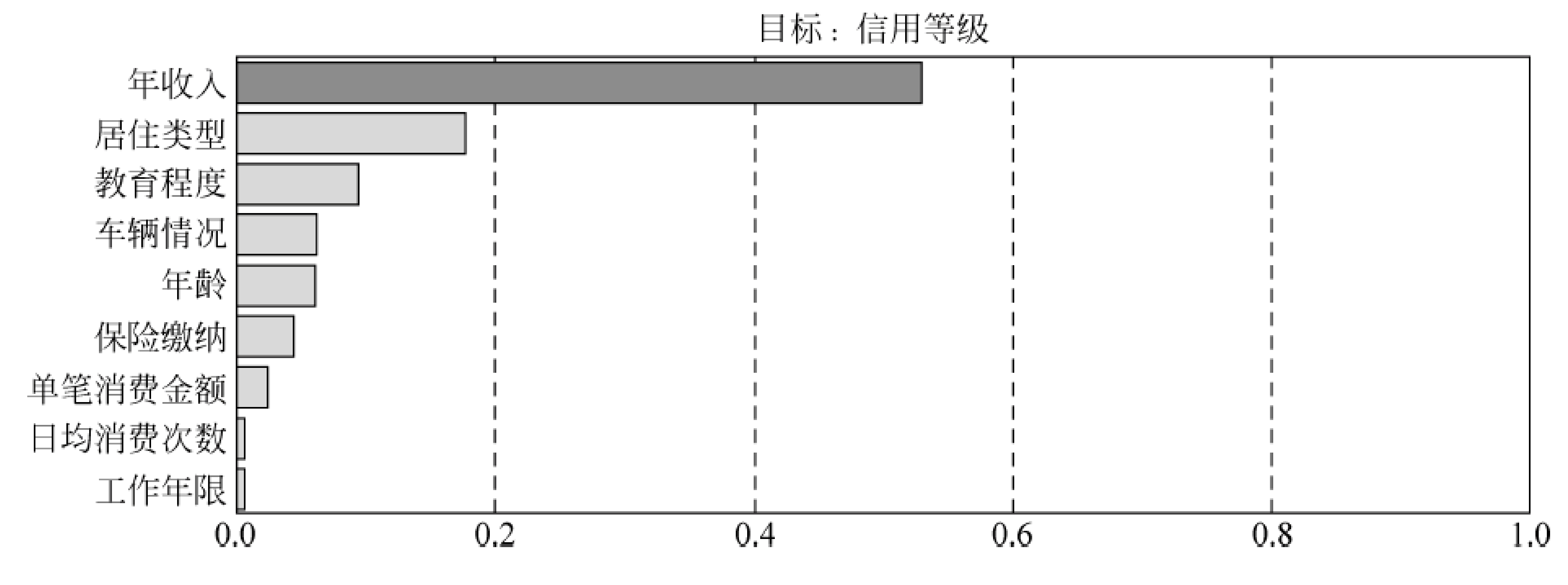


图 6.18 信用等级预测变量重要性分布图

为了进一步分析银行在评判个人用户信用等级时的关注因素，选择合适的决策树层数进行分析，由于得到的决策树共有 9 层，如果全部展开，则得到的决策树不够直观；如果展开层数太少，则不能完整地分析变量重要性，因此需要选择一个合适的决策树层数展开分析。对得到的信用等级影响因素 C5.0 决策树展开 4 层进行分析，得到图 6.19。

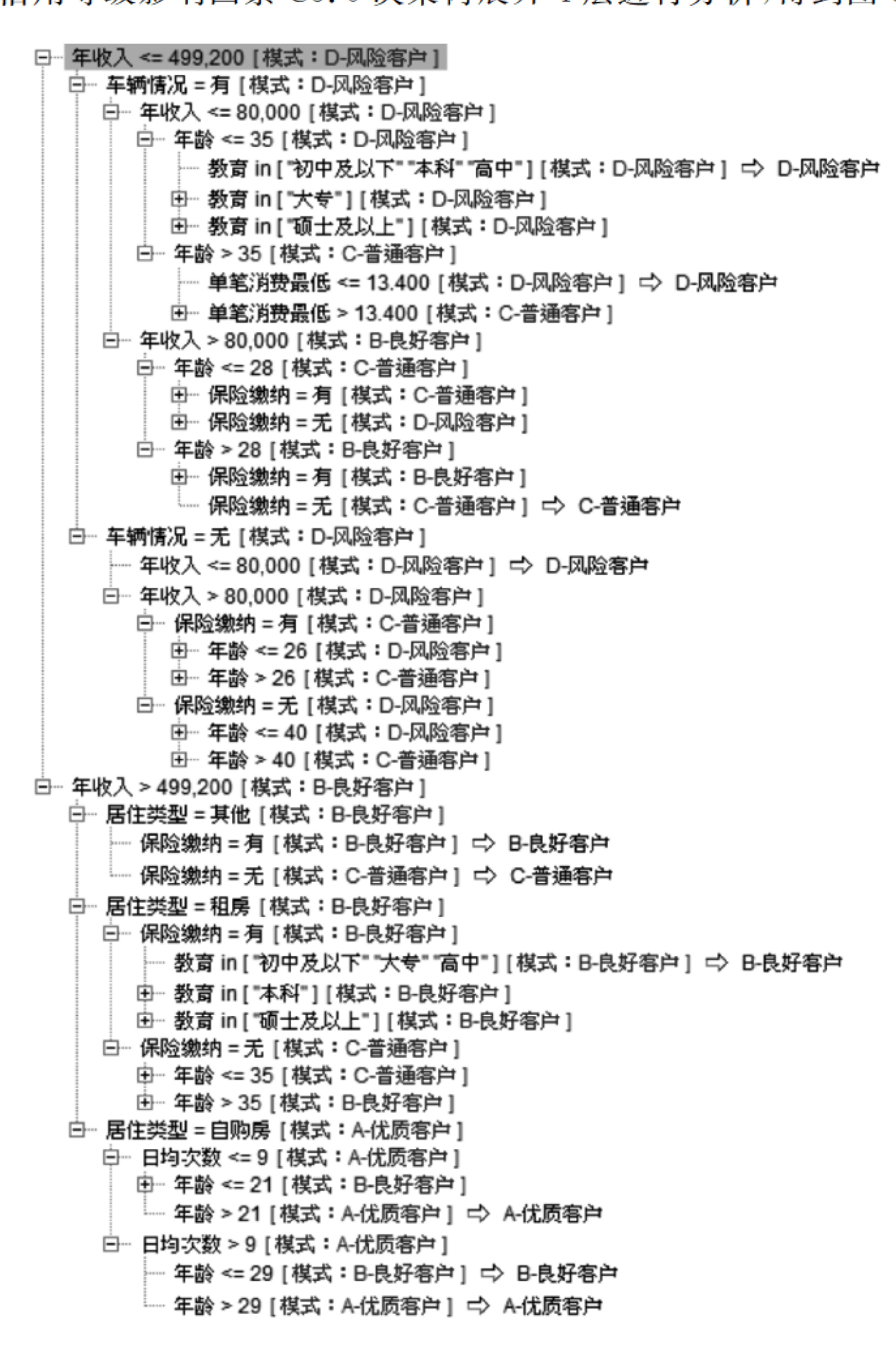


图 6.19　用户信用等级影响因素决策树

可以看到，年收入越高的用户，评价得到的信用等级整体来说就越高，年收入大于 80 000 和年收入小于 80 000 之间的差别最显著；控制年收入不变的情况下，用户的居住类型为自购房，或是有车辆，缴纳了保险，信用等级就越高。年龄在一定情况下，也会影响到个人用户的信用等级。

控制其他变量不变，分析每个变量对用户信用等级影响的原因。

- 用户的年收入越高，用户的消费能力就越强，银行能够从这些用户身上获取的收益就越高。银行信用卡业务的目的就是为银行创造利润，在图 6.19 中可以看到，用户的年收入在 50 万左右时，为优质客户或良好客户。

- 居住类型也体现用户的个人经济实力，为仅次于年收入的重要因素。一般情况下，当用户的居住类型为自购房时，说明经济实力较强。而当用户为租房或是其他居住类型时，说明用户的经济实力较弱或生活不稳定，其信用风险要高于自购房用户，信用等级就低。

通过以上分析，说明个人收入对用户进行信用等级评定时是最重要的。银行信用卡业务的主要目的是盈利，而个人收入较高的用户，能给银行带来的收入就越多。因此，银行在信用卡评级时，主要考虑的因素是用户的个人收入。另一个重要因素是居住类型，它反映的是用户的经济实力，在一定程度上也是个人收入的体现。

银行对用户进行信用等级判定时，应当将个人收入的比重放在第一位，居住类型放在其次位置，着重考虑这两个因素对用户的影响，其他因素作为参考，从而得出银行对用户信用评分时的模型。

6.3 基于消费的信用等级影响因素

信用卡用户的信用等级将随着消费行为的变化而不断调整，调整的依据是消费的行为特征，提供的数据主要为消费历史的统计结果值，如日均消费金额、日均消费次数、单笔消费最小金额、单笔消费最大金额和个人收入。对信用卡用户的消费行为进行统计分析，探寻消费行为与信用等级之间的关系。使用箱图进行分析，日均消费金额的统计结果如图 6.20 所示。

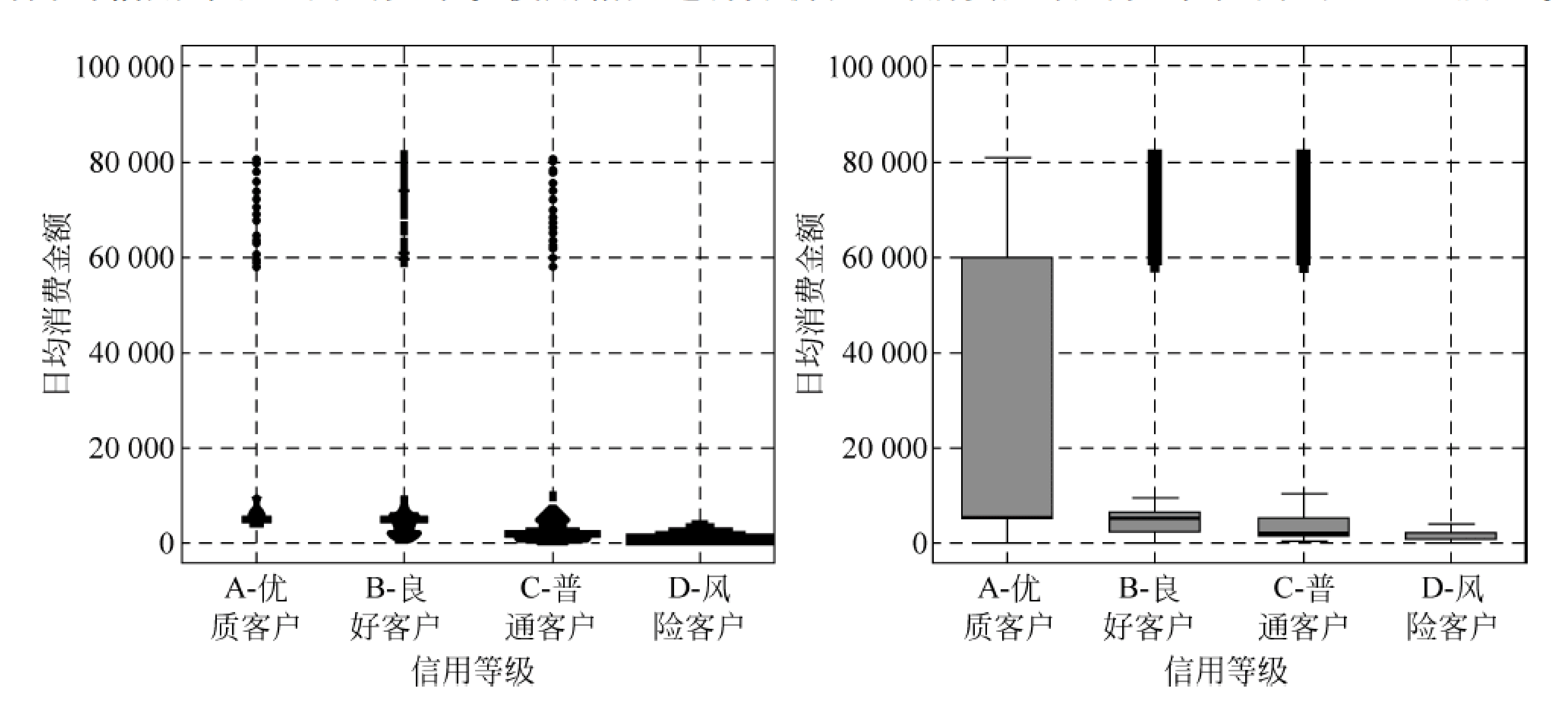


图 6.20 基于日均消费金额的信用等级分析

可以看到，优质客户的日均消费金额比较高，而风险客户普遍较低。用户消费能力越强，其信用卡的等级越高。单笔消费最大金额的箱图如图 6.21 所示，随着客户信用等级的降低，单笔消费最大金额也逐渐减小，特别是在二维点图中，风险客户的最高消费金额基本上集中于 4000 元以下，特征比较明显。

为了量化分析信用等级与消费行为之间的关系，采用 C5.0 算法分析消费行为与信用等级之间的关系，可以发现单笔消费最高具有较大的重要性。如图 6.22 所示，其中单笔消费最高金额低于 33 409 元的客户且其日均消费金额低于 1215 元，为风险客户；日均消费金额高于 1215 元，但单笔消费最高超过了 12 847 元，也为风险客户。

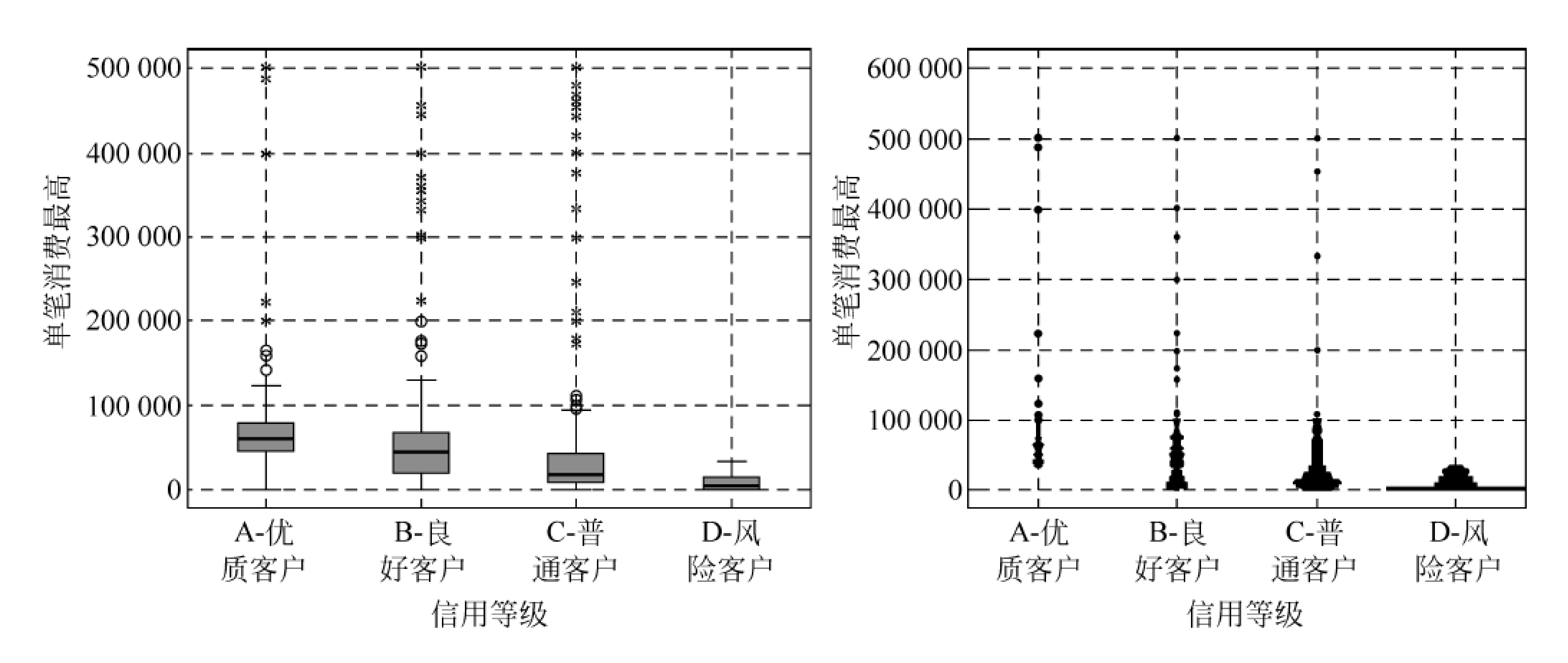


图 6.21　基于单笔消费最大金额的信用等级分析

单笔消费最高 <= 33409.100 [模式：D-风险客户]
日均消费金额 <= 1,215 [模式：D-风险客户] ⇨ D-风险客户
日均消费金额 > 1,215 [模式：C-普通客户]
单笔消费最高 <= 12847.400 [模式：C-普通客户] ⇨ C-普通客户
单笔消费最高 > 12847.400 [模式：D-风险客户] ⇨ D-风险客户
单笔消费最高 > 33409.100 [模式：B-良好客户]
日均消费金额 <= 61,504 [模式：B-良好客户] ⇨ B-良好客户
日均消费金额 > 61,504 [模式：C-普通客户]
日均次数 <= 10 [模式：C-普通客户] ⇨ C-普通客户
日均次数 > 10 [模式：B-良好客户] ⇨ B-良好客户

图 6.22　信用等级与消费行为之间的关系

应用分析节点对分类结果进行评估，在测试集中的准确率只有 53.36%，不具有实际的应用价值，如图 6.23 所示。

输出字段 信用等级 的结果
比较 $C-信用等级 与 信用等级

'分区'	1_培训		2_测试	
正确	1,633	55.89%	1,618	53.36%
错误	1,289	44.11%	1,414	46.64%
总计	2,922		3,032	

图 6.23　基于消费的信用等级分析结果

测试结果较低的原因可能是给定的信用评分为申请信用卡时的评分，并非随着消费行为的变化而改变的动态信用评分，虽然整体分类结果准确率不高，但单笔消费最大金额、日均消费金额较高的用户其消费能力较强(年收入较高)，其相应的信用等级也较高，这与上面客户收入与信用等级呈正相关的结论一致。

6.4　信用卡欺诈判断模型

信用卡欺诈风险是借款人利用信息不对称，骗取信用卡进行恶意透支，严重阻碍了信用卡行业稳健、快速地发展。

随着数据量的快速增长和数据类型日益复杂，信用卡欺诈手段也更加多样化，境外犯罪现象增多，违法分子对商业银行风险核查手段的应变能力增强，信用卡欺诈现象屡禁不止。

2015 年，信用卡欺诈案件数量占经济案件的四分之一，给银行造成经济损失数百亿元，也给银行带来了不可挽回的信誉损失。

6.4.1 基于 Apriori 算法的欺诈模型

通过“消费历史记录”表中的数据，分析用户欺诈行为的发生和消费行为之间的关系。自变量为额度、日均消费金额、日均次数、单笔最大消费金额、个人收入，因变量为是否存在欺诈，见表 6.3。

表 6.3 数据来源与说明

变量类型	变量名	详细说明	取值范围	备注
因变量	是否存在欺诈	定性变量（2 水平）	1 代表存在欺诈；0 代表不存在欺诈	欺诈占比 4.50%
自变量	额度	定性变量（4 水平）	10 000/20 000/50 000/100 000	10 000 占比 39.69%
	日均消费金额	单位：元	30～81 797	只取整数
	日均次数	单位：次	1～28	只取整数
	单笔最大消费金额	单位：元	30.3～500 000	保留一位小数
	个人收入	单位：元	17 000～25 000 000	只取整数

由于日均消费金额、日均次数、单笔最大消费金额、个人收入都是连续变量，不适合使用决策树进行分析，因此衍生出两个新的变量“单笔是否透支”和“日均消费是否超过收入”。“单笔是否透支”根据单笔最大消费金额和信用卡额度得到，若透支，则设为“超过”，否则设为“未超过”。若一个用户单笔消费最大金额－额度>0，则说明该用户的单笔消费存在透支现象，“单笔是否透支”值设为“超过”，如图 6.24 所示。“日均消费是否超过收入”根据用户的年度收入和日均消费金额得到。若一个用户日均消费金额－个人年度收入/365>0，则该用户的日均消费金额超过了收入。若日均消费金额超过了收入，则设为“超过”，否则值设为“未超过”，如图 6.25 所示。



图 6.24 “单笔是否透支”变量定义



图 6.25　“日均消费是否超过收入”变量定义

将刷卡日均次数离散化处理为新的变量“刷卡频率”：1～5 次为不频繁；6～10 次为频繁；11 次及以上为非常频繁，如图 6.26 所示。



图 6.26　“刷卡频率”变量离散化

使用“过滤器”节点将“客户号”“卡号”等标识用户个人的变量过滤，由于“卡类别”与“额度”的作用重复，并且对应关系明确，因此将“卡类别”删除。删除本次分析的无效变量“币种代码”“单笔消费最小金额”，如图 6.27 所示。



图 6.27　“过滤器”属性设置

使用“类型”节点，将“是否存在欺诈”字段设置为目标，上述得到的“单笔是否透支”“日均消费是否超过收入”“刷卡频率”字段设置为输入，使用 Apriori 算法分析用户欺诈行为和消费行为的关系。类型节点属性设置如图 6.28 所示。



图 6.28 “类型”节点属性设置

欺诈判断模型处理流程如图 6.29 所示。

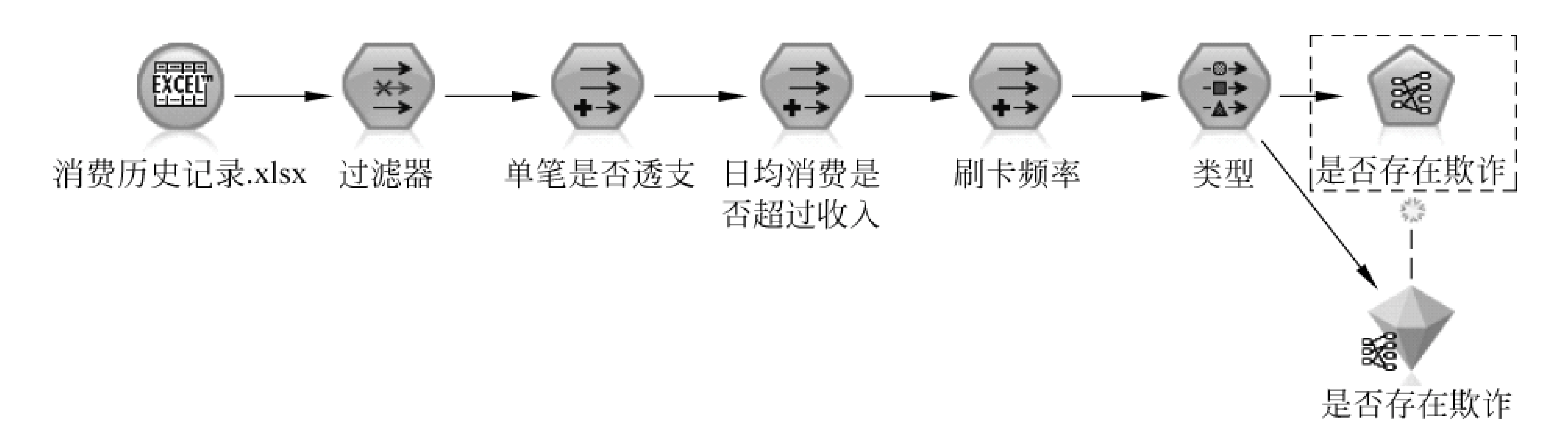


图 6.29 欺诈判断模型处理流程

运行 Apriori 算法，由于欺诈发生所占比例很低，在所有的用户消费记录数据中只占 4.5%，因此最低支持度设置为 0，最低置信度设置为 60%，结果如图 6.30 所示。

后项	前项	支持度百分比	置信度百分比
是否存在欺诈	刷卡频率 = 非常频繁 日均消费是否超过收入	0.286	100.0
是否存在欺诈	刷卡频率 = 非常频繁 单笔是否透支 日均消费是否超过收入	0.151	100.0
是否存在欺诈	刷卡频率 = 频繁 单笔是否透支 日均消费是否超过收入	2.754	100.0
是否存在欺诈	刷卡频率 = 频繁 单笔是否透支	3.829	73.246
是否存在欺诈	刷卡频率 = 非常频繁	0.621	70.27
是否存在欺诈	刷卡频率 = 非常频繁 单笔是否透支	0.453	62.963

图 6.30 欺诈行为与消费记录的关系

可以看出，当用户的刷卡频率为“非常频繁”，即日均次数大于 10 次时，发生欺诈的比例非常高。对于刷卡频率为“非常频繁”和“频繁”的用户，即日均次数大于 5 次时，如果用户同时存在单笔消费透支和日均消费超过收入的情况，则该用户基本存在欺诈行为。但是，由于上述频繁项的支持度百分比数值较低，所以其结果的准确性并不高，为了进一步提升准确

率，可以对原样本进行处理，调整欺诈行为的百分比占比，从而提高频繁项集的支持度百分比比例。

6.4.2　基于判别的欺诈模型

应用判别分析(discriminant analysis)模型进行分析，得出判别函数规则，当有新的记录产生时，可以应用规则判别是否存在欺诈行为。图 6.31 是应用判别模型的流程。

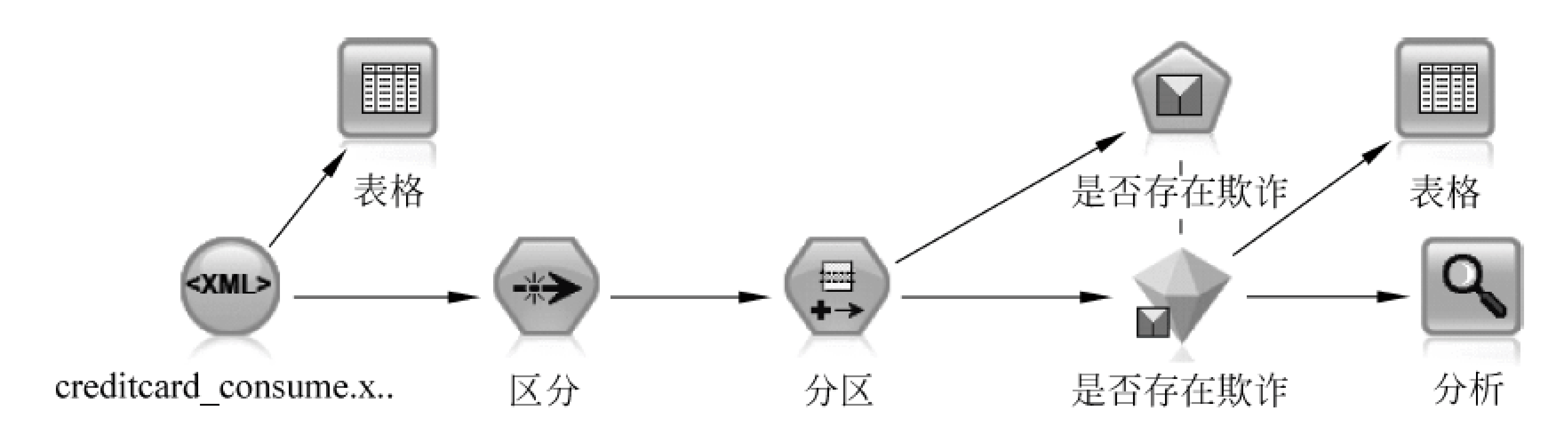


图 6.31　应用判别模型的流程

选择信用记录和消费历史记录，建立训练集 70%和测试集 30%的分区，判别模型的目标字段选为“是否存在欺诈”，输入日均次数、日均消费金额、单笔消费最高、单笔消费最低，并应用分区，如图 6.32 所示。



图 6.32　判别模型字段配置

在“模型”选项卡中选择默认配置，在“专家”选项卡中选择“专家”模式，单击“输出”选项，选择 Box’M、组内相关，在函数系数中选择 Fisher’s。

运行模型后，获得判别模型的结果，其中日均消费次数的权重最高，已远超过 0.8，而日均消费金额、单笔消费最低、单笔消费最高的重要性权重明显偏低。如图 6.33 所示，特征值(Eigenvalues)结果中 Canonical Correlation 表示的是典型相关系数，可以决定变量的相关程度，其中 Wilks’Lambda 的值由 Eigenvalue 计算得出，即 1/(1+0.167)，卡方值为 645.865，

自由度为 2,所以判别函数具有统计上的显著性。

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.167[a]	100.0	100.0	.378

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.857	645.865	2	.000

图 6.33　判别函数显著性检测结果

判别函数的标准化系数表示各自变量与判别函数之间的部分相关系数,即在其他变量不变的情况下,其与目标变量的相关程度,表示自变量的重要程度,从图 6.34 中可以看出日均次数远超过其他变量。

Standardized Canonical Discriminant Function Coefficients

	Function
	1
日均次数	1.009
单笔消费最高	-.175

Structure Matrix

	Function
	1
日均次数	.985
日均消费金额[a]	-.075
单笔消费最低[a]	-.070
单笔消费最高	-.037

图 6.34　判别函数的标准化系数及结构化矩阵系数

结构化矩阵系数表示各自变量与判别函数之间的简单相关程度,与标准化系数相比,结果更加稳定。从图 6.35 中可以看出,其结果与标准化系数相同,日均消费次数重要,其他自变量与目标变量相关性极小。

Classification Function Coefficients

	是否存在欺诈	
	1	0
日均消费次数	1.703	.696
单笔消费最高	2.556E-8	5.592E-6
(Constant)	-6.387	-1.771

Fisher's linear discriminant functions

Classification Results

		是否存在欺诈	Predicted Group Membership 1	Predicted Group Membership 0	Total
Original	Count	1	143	44	187
		0	624	3380	4004
	%	1	76.5	23.5	100.0
		0	15.6	84.4	100.0

图 6.35　分类函数系数及其结果

分类函数系数是基于费雪(R. A. Fisher)的分类函数计算得到的变量系数,通过区分系数的系数值,得到分类的结果,日均消费次数中存在欺诈的系数为 1.703,而无欺诈的系数为 0.696,其他变量的系数差别不大。可以看到,存在欺诈预测的准确率为 76.5%,预测无欺诈行为的准确率为 84.4%,与欺诈识别流程分析节点的结果一致。如图 6.36 所示,训练集的准确率为 84.06%,而测试集的准确率为 83.61%。

6.4.3　基于分类算法的欺诈模型

本节应用 SVM 和 CART 等分类算法对欺诈模型进行分析和构建,如图 6.37 所示,其中输入变量为日均消费次数、日均消费金额、单笔消费最高金额、单笔消费最低金额,目标变

⊟ 输出字段 是否存在欺诈 的结果
⊟ 比较 $D-是否存在欺诈 与 是否存在欺诈

'分区'	1_培训		2_测试	
正确	3,523	84.06%	1,428	83.61%
错误	668	15.94%	280	16.39%
总计	4,191		1,708	

图 6.36 基于判别的预测准确率

量为是否存在欺诈。由于目标变量中的类型分布极不平衡,所以直接应用样本将无法获得应用性较高的模型,需要对样本记录进行平衡,使用"平衡"节点,降低未欺诈记录数为原来的 20%,在分类算法中使用线性 SVM 算法和 CART 模型进行对比分析。

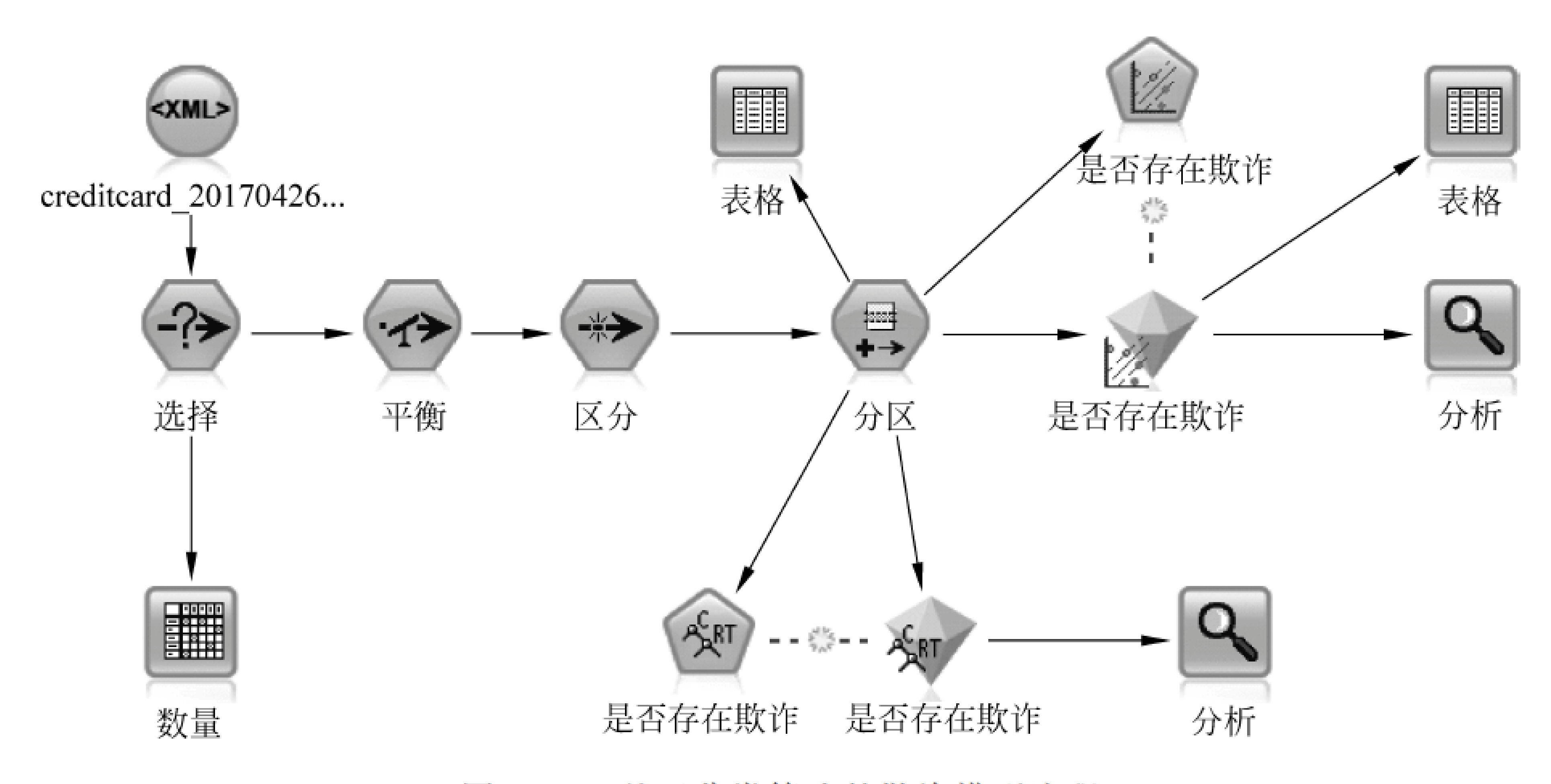


图 6.37 基于分类算法的欺诈模型流程

1. 基于线性 SVM 模型的欺诈模型

分区采用训练集 70%和测试集 30%的比例进行划分,使用"区分"节点将单笔消费最高、单笔消费最低、日均消费金额、日均次数 4 个字段相同的记录滤除重复,经过 LSVM 模型之后,构建了 $L-是否存在欺诈、$LC-是否存在欺诈两列,如图 6.38 所示,其中 $LC-是否存在欺诈表示预测正确的可能性。

	均次数	日均消费金额	卡类别	是否存在欺诈	单笔消费最高	单笔消费最低	分区	$L-是否存在欺诈	$LC-是否存在欺诈
1	1	30	白金卡	0	30.300	1.500	2_测试	0	0.762
2	3	30	金卡	0	30.300	1.500	2_测试	0	0.719
3	2	31	金卡	0	31.000	1.500	2_测试	0	0.741
4	5	31	金卡	0	30.400	1.500	2_测试	0	0.670
5	7	31	金卡	0	30.600	1.500	1_培训	0	0.618
6	2	39	普卡	0	36.000	3.000	2_测试	0	0.741
7	2	42	普卡	0	40.000	3.000	1_培训	0	0.741
8	2	46	普卡	0	43.000	3.000	2_测试	0	0.741
9	7	48	普卡	0	45.500	3.300	1_培训	0	0.618
10	2	53	普卡	0	50.200	3.900	1_培训	0	0.741
11	6	55	普卡	0	52.000	4.000	2_测试	0	0.645
12	5	57	普卡	0	55.300	4.400	1_培训	0	0.670

图 6.38 基于线性 SVM 算法的模型结果列表

模型的信息和混淆矩阵的信息如图 6.39 所示,分类的准确性为 89.2%,其中 1 表示客户欺诈,0 表示没有欺诈行为,预测存在欺诈且成功的概率为 58%,预测未欺诈且成功的概率为 96%。

模型信息

目标字段	是否存在欺诈
模型构建方法	线性 SVM
输入的预测变量数	4
最终模型中的预测变量数	4
规则化类型	L2
惩罚参数 (Lambda)	0.100
分类准确性	89.2%

混淆矩阵

实测	预测		
	1	0	比例正确
1	107	79	0.58
0	29	789	0.96
比例正确	0.79	0.91	0.89

图 6.39 模型的信息和混淆矩阵的信息

分析发现，日均次数变量的重要性权重最高，超过 0.75，单笔消费最高金额、单笔消费最低金额次之，日均消费金额最不重要。

使用箱图分析是否有欺诈行为的日均消费次数，其中 0 表示没有欺诈，1 表示有欺诈行为。可以看到，有欺诈行为的用户日均消费次数明显偏多，如图 6.40 所示。

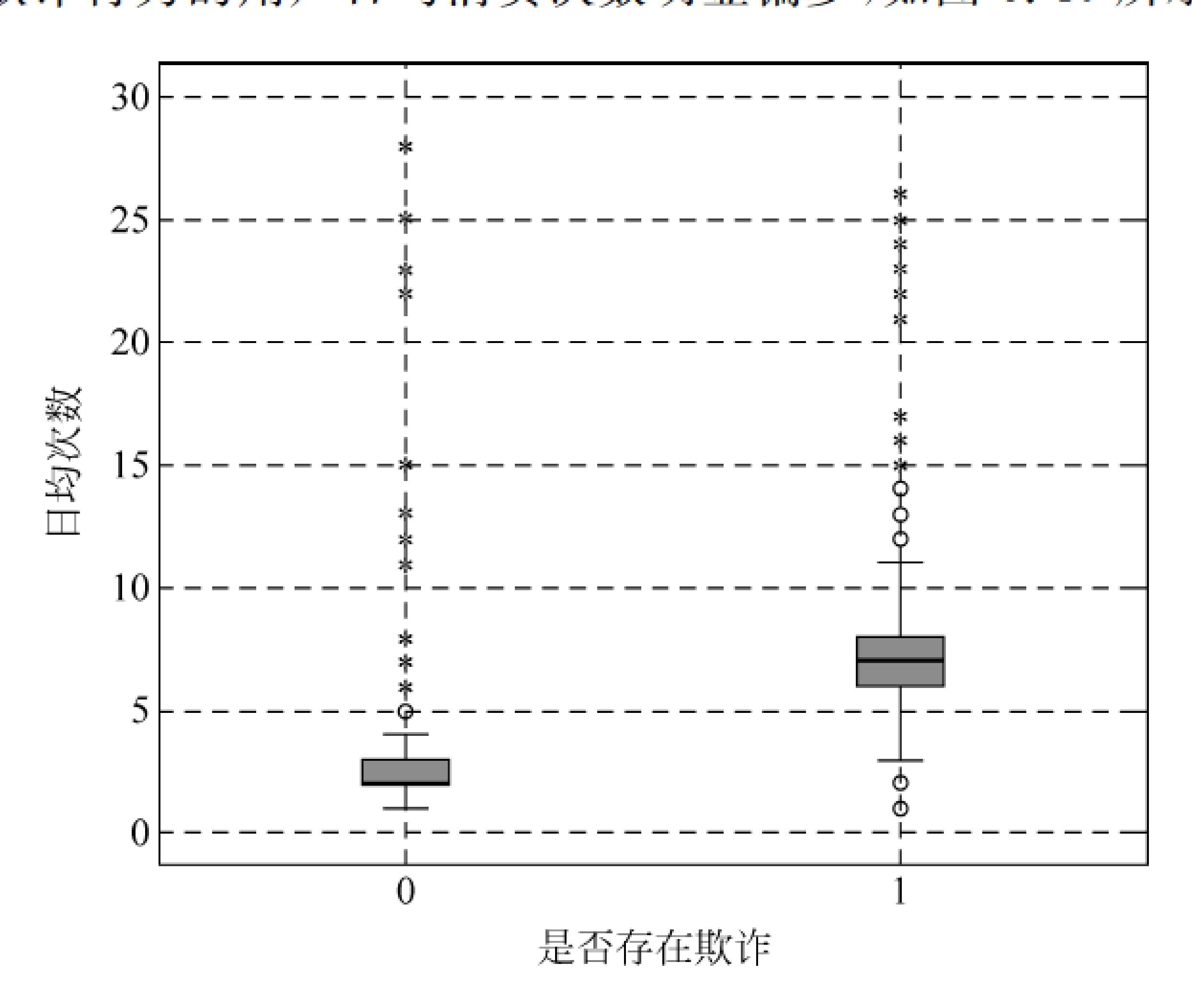


图 6.40 信用卡欺诈与日均消费次数的关系

应用分析节点查看模型中训练集和测试集的准确率，可以看到在测试集中达到 88.06% 的准确率，预测错误的记录只有 45 条，占总数的 11.94%，说明模型具有一定的应用价值，如图 6.41 所示。

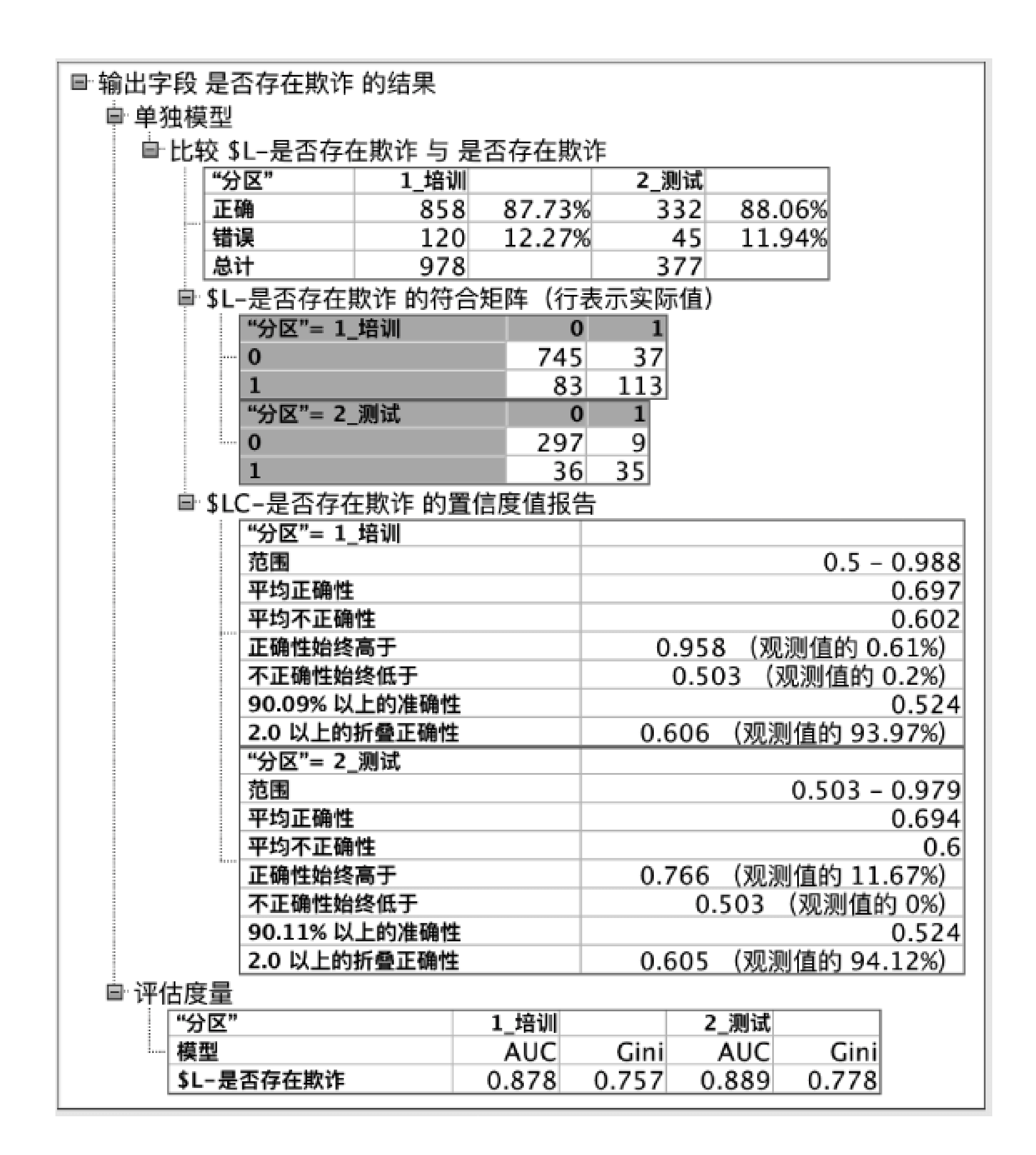
虽然模型的整体指标准确率较高，但是从真阳率的指标来看，其预测为欺诈的准确率仅为 58%，在实际应用中效果可能并不理想。

2. 基于 CART 模型的欺诈模型

使用 CART 进行对比分析，结果如图 6.42 所示。

分析发现，其中日均次数比重较高，单笔消费最低金额最低、单笔消费最高金额次之，最不重要的是日均消费金额。

由 CART 算法得到决策树如图 6.43 所示，其中日均次数超过 5.5 次，单笔消费最高金额大于 9558.2 元，可标记为具有欺诈行为。



输出字段 是否存在欺诈 的结果

单独模型

比较 $L-是否存在欺诈 与 是否存在欺诈

"分区"	1_培训		2_测试	
正确	858	87.73%	332	88.06%
错误	120	12.27%	45	11.94%
总计	978		377	

$L-是否存在欺诈 的符合矩阵（行表示实际值）

"分区"= 1_培训	0	1
0	745	37
1	83	113

"分区"= 2_测试	0	1
0	297	9
1	36	35

$LC-是否存在欺诈 的置信度值报告

"分区"= 1_培训	
范围	0.5 - 0.988
平均正确性	0.697
平均不正确性	0.602
正确性始终高于	0.958（观测值的 0.61%）
不正确性始终低于	0.503（观测值的 0.2%）
90.09% 以上的准确性	0.524
2.0 以上的折叠正确性	0.606（观测值的 93.97%）

"分区"= 2_测试	
范围	0.503 - 0.979
平均正确性	0.694
平均不正确性	0.6
正确性始终高于	0.766（观测值的 11.67%）
不正确性始终低于	0.503（观测值的 0%）
90.11% 以上的准确性	0.524
2.0 以上的折叠正确性	0.605（观测值的 94.12%）

评估度量

"分区"	1_培训		2_测试	
模型	AUC	Gini	AUC	Gini
$L-是否存在欺诈	0.878	0.757	0.889	0.778

图 6.41　基于线性 SVM 算法的模型结果分析

	信用等级	总评分	额度	日均次数	日均消费金额	卡类别	是否存在欺诈	单笔消费最高	单笔消费最低	分区	$R-是否存在欺诈	$RC-是否存在欺诈
1	D-风险客户	60	10000	4	106	普卡	0	129.900	6.600	1_培训	0	0.989
2	D-风险客户	60	10000	3	101	普卡	0	102.000	5.000	1_培训	0	0.989
3	D-风险客户	60	10000	4	106	普卡	0	130.200	6.700	2_测试	0	0.989
4	D-风险客户	60	10000	2	107	普卡	0	131.500	6.900	2_测试	0	0.989
5	D-风险客户	60	10000	2	107	普卡	0	131.800	6.900	1_培训	0	0.989
6	D-风险客户	60	10000	5	107	普卡	0	134.100	7.000	1_培训	0	0.989
7	D-风险客户	60	10000	1	101	普卡	0	102.100	5.000	2_测试	0	0.989
8	D-风险客户	60	10000	7	107	普卡	0	134.100	7.000	1_培训	0	0.974
9	D-风险客户	60	10000	1	101	普卡	0	104.800	5.000	1_培训	0	0.989
10	D-风险客户	60	10000	4	108	普卡	0	140.000	7.300	1_培训	0	0.989
11	D-风险客户	60	10000	2	109	普卡	0	141.200	7.500	2_测试	0	0.989
12	D-风险客户	60	10000	2	109	普卡	0	141.600	7.500	2_测试	0	0.989
13	D-风险客户	60	10000	2	102	普卡	0	108.800	5.000	2_测试	0	0.989
14	D-风险客户	60	10000	2	102	普卡	0	109.400	5.000	1_培训	0	0.989
15	D-风险客户	60	10000	5	103	普卡	0	114.700	5.500	2_测试	0	0.989
16	D-风险客户	60	10000	2	104	普卡	0	120.900	5.900	2_测试	0	0.989
17	D-风险客户	60	10000	5	111	普卡	0	149.000	7.800	1_培训	0	0.989
18	D-风险客户	60	10000	2	113	普卡	0	151.300	7.800	1_培训	0	0.989
19	D-风险客户	60	10000	1	114	普卡	0	153.900	7.900	2_测试	0	0.989
20	D-风险客户	60	10000	2	116	普卡	0	156.600	8.000	1_培训	0	0.989
21	D-风险客户	60	10000	2	117	普卡	0	159.600	8.100	1_培训	0	0.989
22	D-风险客户	60	10000	4	119	普卡	0	160.200	8.200	2_测试	0	0.989

图 6.42　CART 算法分析结果

使用分析节点，查看 CART 的模型分析结果，如图 6.44 所示。可以看到，这个模型在训练集和测试集中均有较好的表现，达到 95.66％的准确率，从符合矩阵中可以计算得到其真阳率达到 80.25％，具有一定的应用价值。

因此，银行在判断用户是否存在欺诈行为时，可以从用户的消费记录着手，关注用户的刷卡频率，并且对用户“单笔消费是否透支”以及“日均消费是否超过收入”进行记录，从而及早发现可能有欺诈行为发生，对于很有可能产生欺诈行为的用户，及时采取预警，避免用户继续进行欺诈行为，从而减少欺诈行为给银行带来的经济损失。

```
日均次数 <= 5.500 [ 模式: 0 ] ⇨ 0
日均次数 > 5.500 [ 模式: 1 ]
    单笔消费最高 <= 9558.200 [ 模式: 0 ] ⇨ 0
    单笔消费最高 > 9558.200 [ 模式: 1 ] ⇨ 1
```

图 6.43　CART 决策树模型

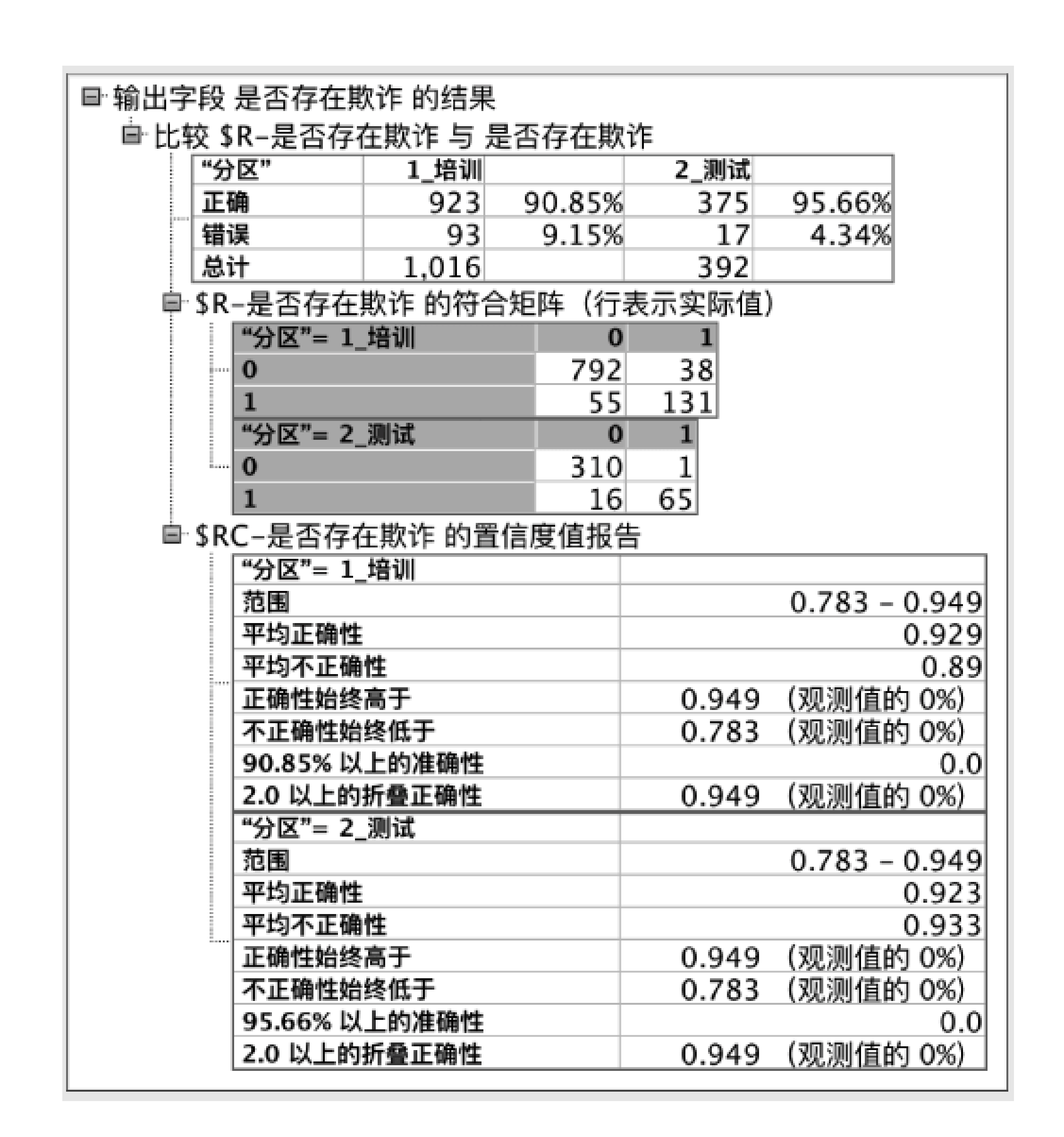


图 6.44　CART 决策树模型结果

6.5　欺诈人口属性分析

在分析欺诈模型的基础上，为了进一步分析何种用户容易发生欺诈行为，对用户的人口属性变量进行统计分析，选择与用户人口属性有关的字段，这些字段统称为客户因素。数据类型与说明见表 6.4。

表 6.4　数据来源与说明

变量类型	变量名	详 细 说 明	取 值 范 围	备　注
因变量	是否存在欺诈	定性变量（2 水平）	1 代表存在欺诈；0 代表不存在欺诈	欺诈占比 4.50%
自变量：客户因素	性别	定性变量（2 水平）	男、女	男性占比 71.01%
	年龄	单位：岁	18～80	只取整数
	婚姻状况	定性变量（4 水平）	离异/丧偶/未婚/已婚	未婚占比 65.12%
	户籍	定性变量（30 水平）	全国各省	
	教育程度	定性变量（5 水平）	初中及以下/高中/大专/本科/硕士及以上	本科占比 49.36%
	居住类型	定性变量（3 水平）	租房/自购房/其他	租房占比 68.74%
	职业类型	定性变量（5 水平）	个体户/国有企业/私营企业/外资企业/其他企业	私营企业占比 59.71%
	工作年限	单位：年	0～50	只取整数
	个人收入	单位：元	10 416～99 000 000 000	只取整数
	保险缴纳	定性变量（2 水平）	有/无	有占比 66.75%
	车辆情况	定性变量（2 水平）	有/无	无占比 65.85%

6.5.1　欺诈人口属性统计分析

在 Excel 中，将“消费历史记录”和“客户信用记录”两个表按照关键词“客户号”进行合并，删除“日均消费金额”“日均次数”“单笔消费最小金额”“单笔消费最大金额”等不需要的字段，得到一个新表。用户人口属性信息与欺诈的关系，如图 6.45 所示。

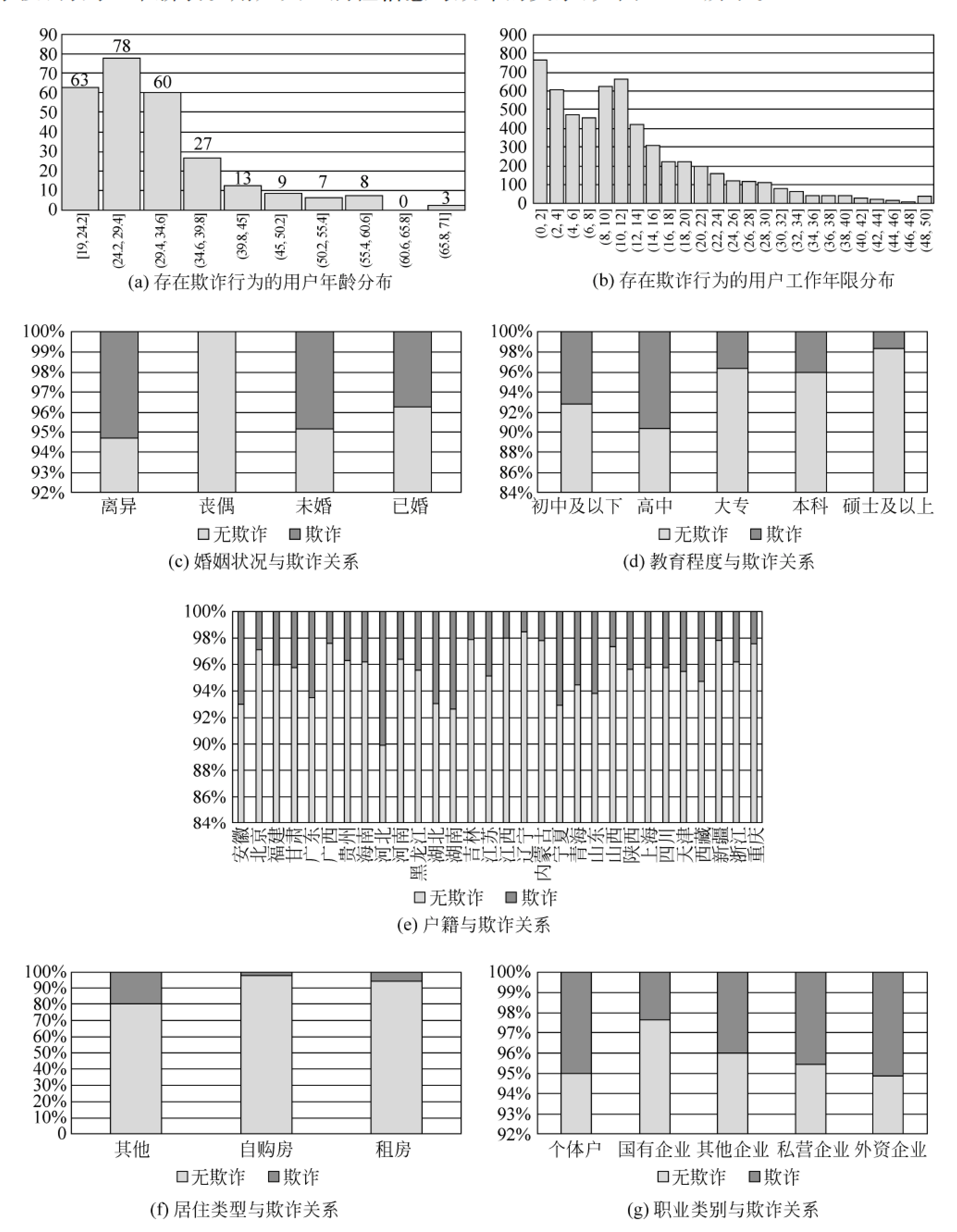
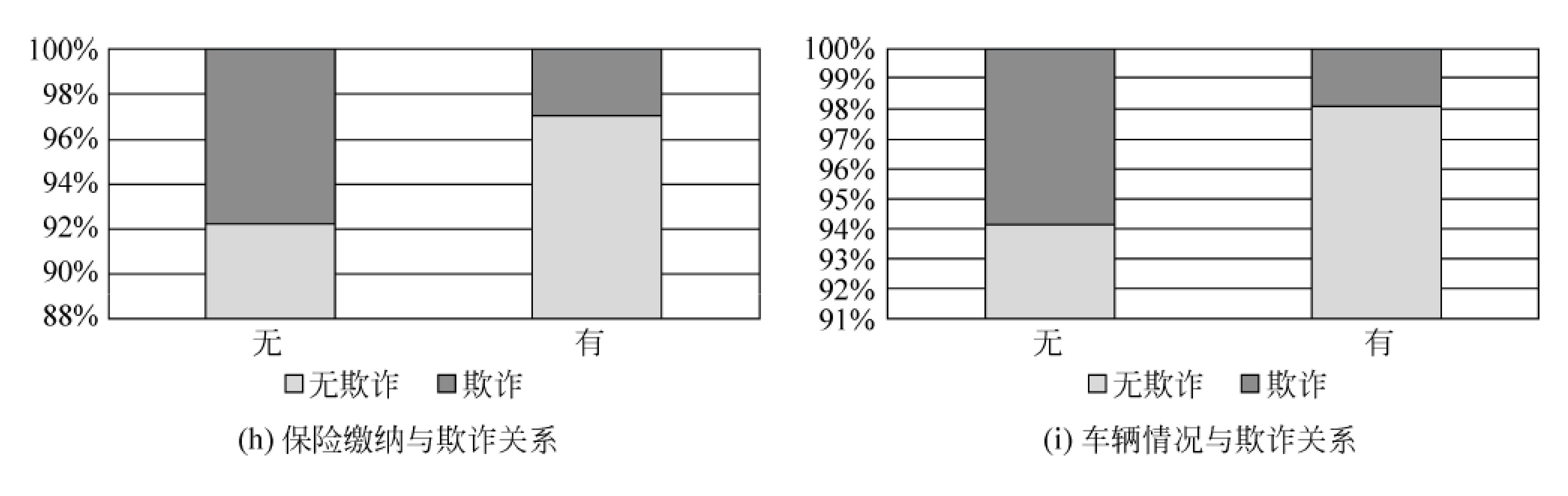


图 6.45　用户人口属性信息与欺诈的关系



(h) 保险缴纳与欺诈关系　　(i) 车辆情况与欺诈关系

图 6.45 （续）

整体来看，用户的工作、生活越稳定，受教育水平和收入水平越高，发生欺诈的比例就越低。

6.5.2 基于逻辑回归的欺诈人口属性分析

为了更加深入地了解用户信用卡欺诈行为的发生原因及其相对重要性，可以对用户记录进行回归分析。在 SPSS Modeler 18.0 中，合并“客户信用记录”和“消费历史记录”两个表。

使用“过滤器”节点，将“客户号”“客户姓名”等标识用户个人的变量过滤。删除无效变量“币种代码”“日均消费金额”“日均次数”“单笔消费最小金额”“单笔消费最大金额”等字段，只剩下与用户人口属性有关的字段，如图 6.46 所示。

过滤器 注解

字段：已输入 25 个，已过滤 13 个，已重命名 0 个，已输出 12 个

字段	过滤器	字段
客户号	—✕→	客户号
客户姓名	—✕→	客户姓名
性别	⟶	性别
年龄_连续	⟶	年龄_连续
婚姻状态	⟶	婚姻状态
户籍	⟶	户籍
教育程度	⟶	教育程度
居住类型	⟶	居住类型
职业类别	⟶	职业类别
工作年限	⟶	工作年限

◉ 查看当前字段 ○ 查看未使用的字段设置

图 6.46 “过滤器”节点属性设置

使用“类型”节点，将“是否存在欺诈”字段设置为目标，“户籍”“教育程度”“居住类型”“职业类别”等人口属性字段设置为输入，使用逻辑回归算法分析用户欺诈行为和消费行为的关系，如图 6.47 所示。

逻辑回归主要在流行病学中应用较多，比较常用的情形是探索某疾病的危险因素，根据危险因素预测某疾病发生的概率。而信用卡欺诈行为也可以看成是一种类似疾病的不良结果，欺诈行为的发生类似于疾病的发生，而用户的个人信息，用户的消费行为作为诱发这种不良结果的危险因素，因此采用逻辑回归，寻找导致信用卡欺诈行为发生的危险因素，并且通过得到的模型，预测在不同危险因素变量值的情况下，用户发生信用卡欺诈行为的可能性，如图 6.48 所示。



图 6.47 “类型”节点属性设置

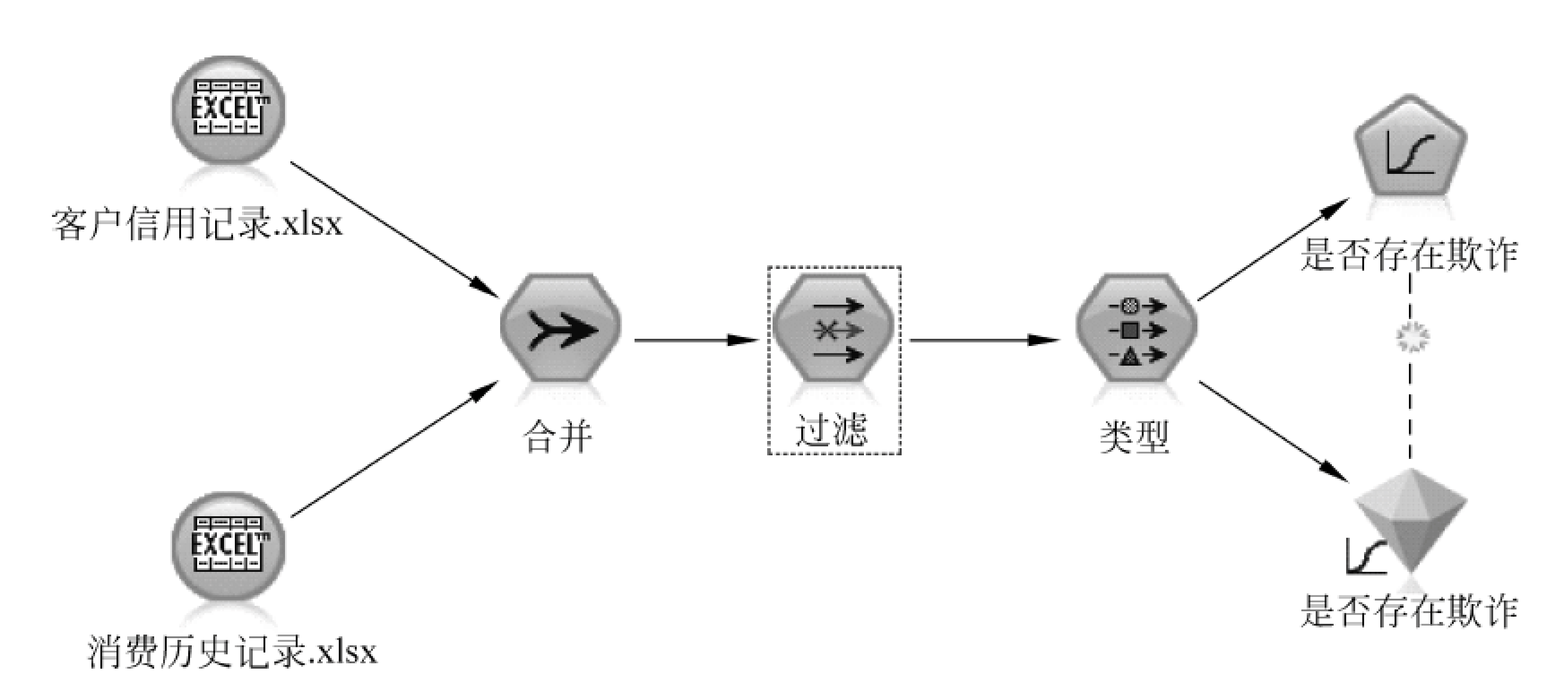


图 6.48 欺诈人口属性分析流程

逻辑模型的检验结果如图 6.49 所示,显示了模型的拟合效果。图 6.50 显示了欺诈逻辑回归分析结果。

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	2187.797	2194.489	2185.797			
Final	2047.309	2375.208	1949.309	236.489	48	.000

图 6.49 逻辑模型的检验结果

从图 6.50 可以看到,与欺诈相关的主要因素包括年收入、年龄、户籍(安徽、河北、湖北、湖南、宁夏)、居住类型(其他)、车辆情况(无)、保险缴纳(无)、教育(本科、高中),分析结果与上一节的分析结果一致。

因此,如果一个用户为高中学历的未婚人士,没有固定的住所,在其他类别的企业工作,没有私人车辆和保险缴纳,那么这个用户发生信用卡欺诈的概率就要远远高于其他用户,银行可以降低其信用额度,提早做好风险防控。

对于欺诈行为地域性差异明显的结果,银行可以调整旗下各地支行的营销策略,对于欺诈行为容易发生的地方,提高申请信用卡的门槛,提高管理费用和服务费用。对于欺诈风险低的地方,降低管理费用,适当降低申请信用卡的要求,让欺诈风险低的地方有更多用户能够享受到信用卡服务。在营销宣传的时候,可以采取地区差异性宣传的方式,针对各地用户不同的整体信用水平,调整银行在各地的业务类别和业务内容。

Parameter Estimates

是否存在欺诈[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
0	Intercept	5.157	.865	35.573	1	.000			
	工作年限	-.013	.013	1.001	1	.317	.987	.963	1.012
	年收入	.000	.000	12.199	1	.000	1.000	1.000	1.000
	年龄	.031	.013	5.392	1	.020	1.031	1.005	1.058
	[户籍=安徽]	-1.192	.539	4.879	1	.027	.304	.105	.874
	[户籍=北京]	-.250	.585	.182	1	.670	.779	.247	2.453
	[户籍=福建]	-.475	.608	.610	1	.435	.622	.189	2.048
	[户籍=甘肃]	-.669	.603	1.229	1	.268	.512	.157	1.671
	[户籍=广东]	-.951	.506	3.537	1	.060	.386	.143	1.041
	[户籍=广西]	.164	.691	.057	1	.812	1.179	.304	4.569
	[户籍=贵州]	-.488	.626	.608	1	.435	.614	.180	2.093
	[户籍=海南]	-.486	.604	.647	1	.421	.615	.188	2.009
	[户籍=河北]	-1.595	.512	9.694	1	.002	.203	.074	.554
	[户籍=河南]	-.493	.603	.669	1	.413	.611	.187	1.991
	[户籍=黑龙江]	-.646	.588	1.204	1	.273	.524	.166	1.662
	[户籍=湖北]	-1.216	.540	5.076	1	.024	.296	.103	.854
	[户籍=湖南]	-1.237	.540	5.246	1	.022	.290	.101	.837
	[户籍=吉林]	.107	.686	.024	1	.876	1.113	.290	4.270
	[户籍=江苏]	-.831	.588	1.999	1	.157	.436	.138	1.379
	[户籍=江西]	.189	.685	.076	1	.783	1.208	.315	4.623
	[户籍=辽宁]	.473	.745	.403	1	.526	1.604	.373	6.908
	[户籍=内蒙古]	.229	.689	.110	1	.740	1.257	.326	4.851
	[户籍=宁夏]	-1.262	.546	5.350	1	.021	.283	.097	.825
	[户籍=青海]	-.940	.558	2.845	1	.092	.390	.131	1.165
	[户籍=山东]	-.912	.561	2.641	1	.104	.402	.134	1.207
	[户籍=山西]	-.107	.648	.027	1	.868	.898	.252	3.199
	[户籍=陕西]	-.801	.576	1.932	1	.165	.449	.145	1.389
	[户籍=上海]	-.638	.520	1.509	1	.219	.528	.191	1.463
	[户籍=四川]	-.701	.575	1.488	1	.223	.496	.161	1.530
	[户籍=天津]	-.709	.575	1.519	1	.218	.492	.159	1.520
	[户籍=西藏]	-.842	.869	.939	1	.332	.431	.078	2.365
	[户籍=新疆]	-.185	1.119	.027	1	.869	.831	.093	7.451
	[户籍=浙江]	-.617	.587	1.105	1	.293	.540	.171	1.705
	[户籍=重庆]	0[b]	.	.	0	.	.	.	.
	[居住类型=其他]	-1.567	.349	20.192	1	.000	.209	.105	.413
	[居住类型=自购房]	-.405	.549	.545	1	.460	.667	.227	1.955
	[居住类型=租房]	0[b]	.	.	0	.	.	.	.
	[车辆情况=无]	-1.035	.517	4.009	1	.045	.355	.129	.978
	[车辆情况=有]	0[b]	.	.	0	.	.	.	.
	[保险缴纳=无]	-.936	.205	20.807	1	.000	.392	.262	.586
	[保险缴纳=有]	0[b]	.	.	0	.	.	.	.
	[性别=男]	-.004	.145	.001	1	.980	.996	.750	1.323
	[性别=女]	0[b]	.	.	0	.	.	.	.
	[婚姻=离异]	-.352	.303	1.350	1	.245	.703	.388	1.274
	[婚姻=丧偶]	17.262	.000	.	1	.	31391466.64	31391466.64	31391466.64
	[婚姻=未婚]	-.201	.152	1.750	1	.186	.818	.607	1.102
	[婚姻=已婚]	0[b]	.	.	0	.	.	.	.
	[教育=本科]	-1.140	.347	10.823	1	.001	.320	.162	.631
	[教育=初中及以下]	-.787	.447	3.100	1	.078	.455	.189	1.093
	[教育=大专]	-.334	.379	.778	1	.378	.716	.341	1.504
	[教育=高中]	-1.386	.390	12.657	1	.000	.250	.117	.537
	[教育=硕士及以上]	0[b]	.	.	0	.	.	.	.
	[职业=个体户]	-.011	.249	.002	1	.966	.989	.608	1.611
	[职业=国有企业]	.566	.386	2.152	1	.142	1.762	.827	3.754
	[职业=其他企业]	-.250	.320	.612	1	.434	.779	.416	1.457
	[职业=私营企业]	.012	.192	.004	1	.952	1.012	.695	1.473
	[职业=外资企业]	0[b]	.	.	0	.	.	.	.

a. The reference category is: 1.

b. This parameter is set to zero because it is redundant.

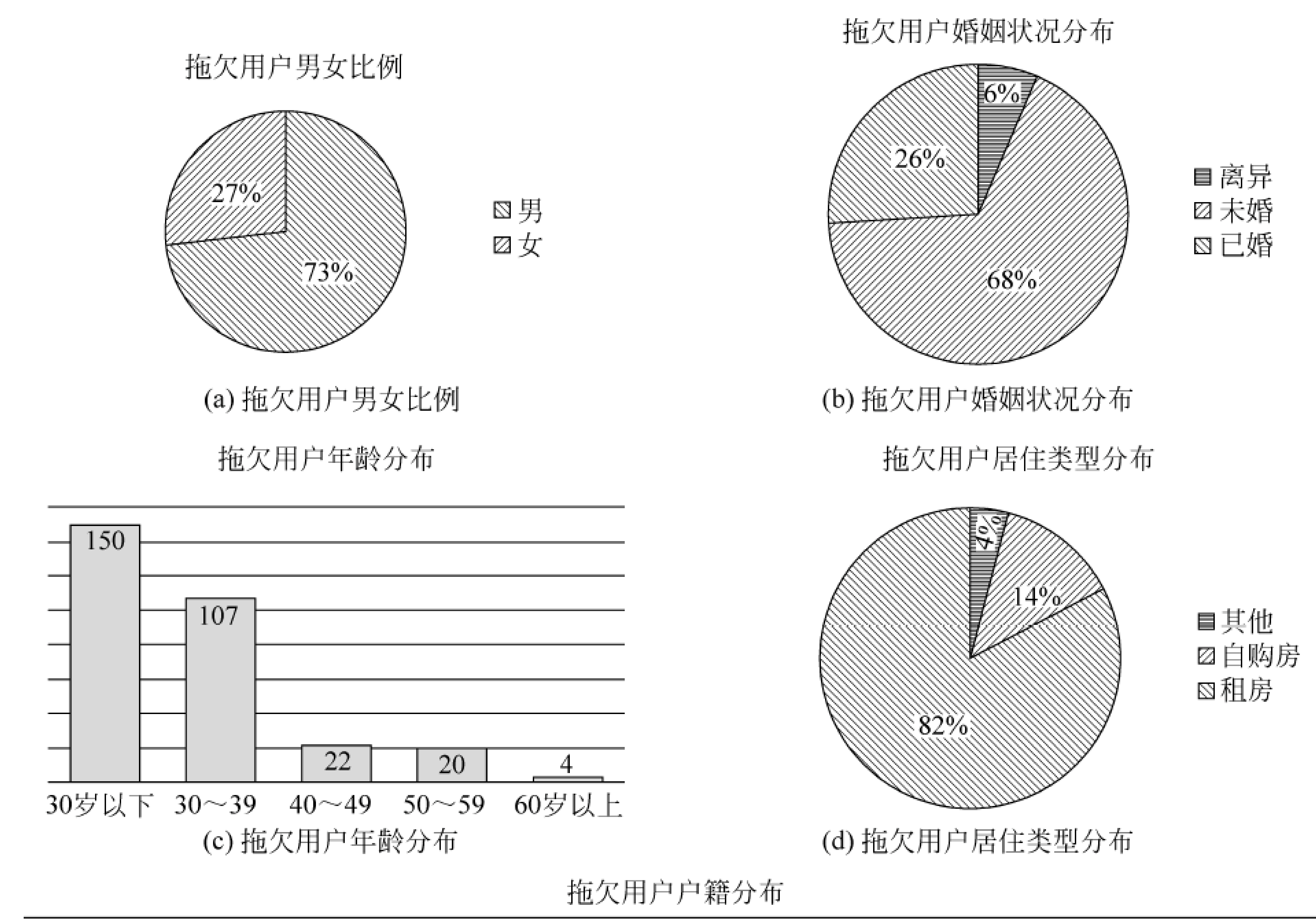
图 6.50 欺诈逻辑回归分析结果

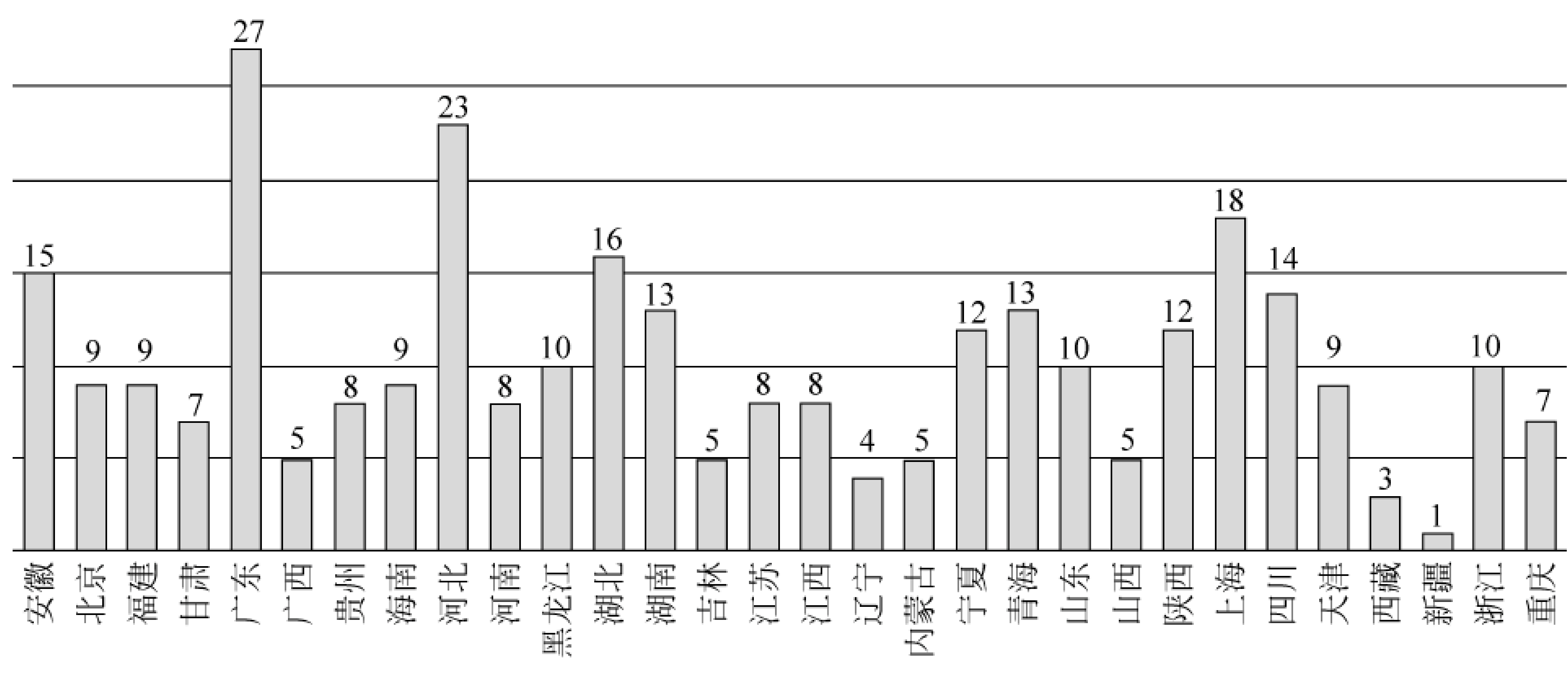
6.5.3　逾期还款的客户特征

信用卡拖欠与欺诈为银行信用卡业务非人为操作的两大风险，会给银行带来巨大的经济损失。信用风险是指借款人不能在规定期限内按照约定的合约及时、足额偿还银行本金和利息的可能性。银行需要及早根据用户的个人信息，评估用户发生拖欠行为的可能性，通过减少用户借贷额度等行为尽早做好风险防控工作。

下面通过银行的客户信息和拖欠历史记录，对客户的个人信息进行分析，从而对产生拖欠的用户进行画像，得到容易发生拖欠的用户模型，为银行的风险管控工作提供参考，从而降低银行的损失。

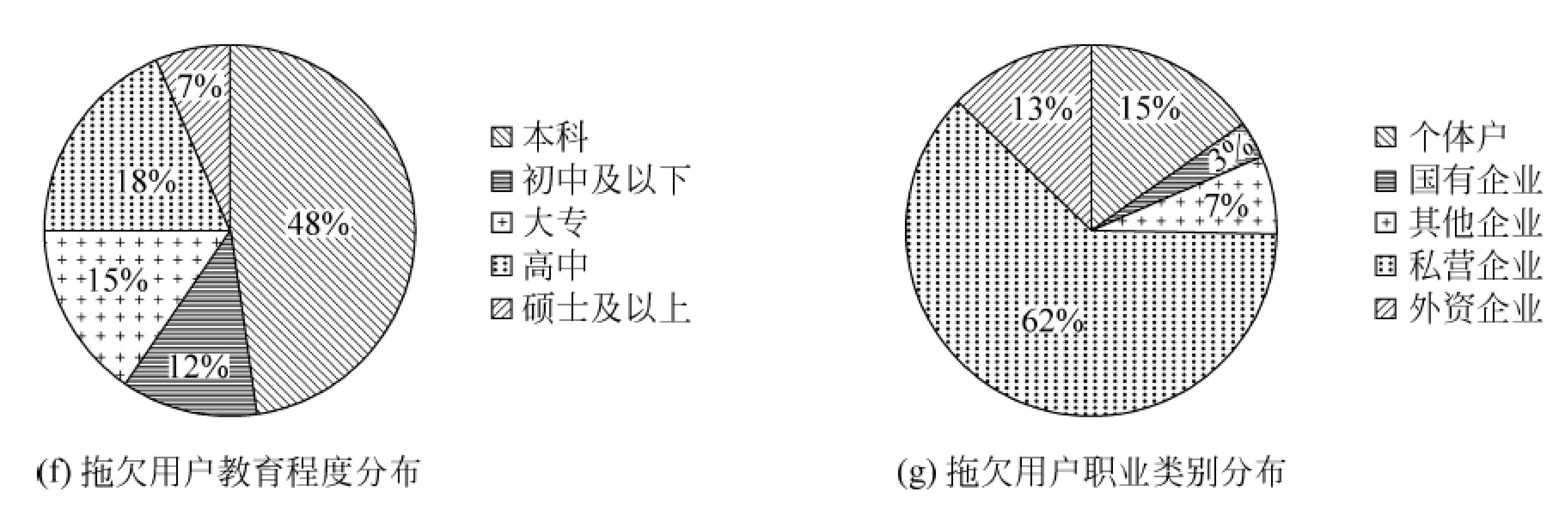
使用 C5.0 算法，分析客户信用记录和拖欠历史记录两张表，找出逾期客户的画像，如图 6.51 所示。





(e) 拖欠用户户籍分布

图 6.51　拖欠用户按人口属性分布



(f) 拖欠用户教育程度分布　　(g) 拖欠用户职业类别分布

图 6.51 （续）

可以看出，拖欠用户具有以下特征：男性、未婚用户、40 岁以下、租房、本科、私营企业工作，拖欠用户在广东、河北、上海、安徽、湖北、四川人数较多。

6.5.4 基于决策树分析逾期客户特征

首先需要对客户的拖欠程度进行评估，拖欠历史记录表中有两个字段与拖欠程度评估相关，一个为“拖欠总金额”，另一个为“逾期天数”，可以将其结合起来用一系列步骤得到拖欠程度的计量化评估。对“拖欠金额”进行打分评估得到“拖欠金额得分”，如图 6.52 所示。

公式:

```
if (拖欠总金额 <= 2500) then 20
  else if(拖欠总金额 <= 5000) then 60
    else if(拖欠总金额 <= 22000) then 75
      else 100
    endif
  endif
endif
```

图 6.52 拖欠金额得分

对“逾期天数”进行打分评估得到“拖欠时间得分”，如图 6.53 所示。

公式:

```
if (逾期天数 <= 30) then 20
  else if (逾期天数 <= 60) then 60
    else if (逾期天数 <= 90) then 75
      else 100
    endif
  endif
endif
```

图 6.53 拖欠时间得分

根据“拖欠金额得分”和“拖欠时间得分”按照一定的比例得到“拖欠总得分”为拖欠金额得分×0.6＋拖欠时间得分×0.4。根据“拖欠总得分”得到拖欠程度划分的公式如图 6.54 所示。

将上述过程结合起来可以得到对拖欠历史表处理获取拖欠程度的数据挖掘流，如图 6.55 所示。

公式:

```
if (拖欠总得分 <= 60) then '轻度拖欠'
  else if (拖欠总得分 <= 75) then '中度拖欠'
    else '重度拖欠'
  endif
endif
```

图 6.54 拖欠总得分离散划分

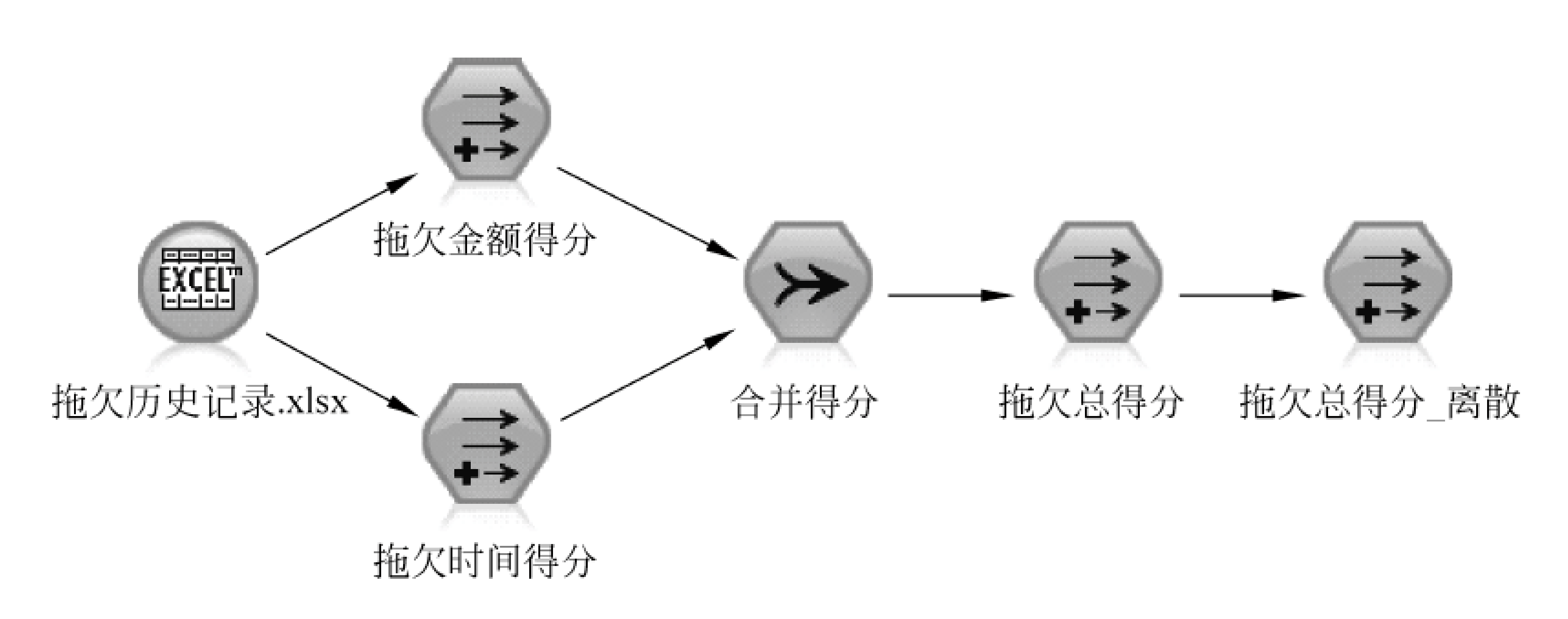


图 6.55 获取拖欠程度的数据挖掘流

然后对客户信用记录表进行初步处理,将“居住类型”中的“自购房”等同于“有房”“租房”和“其他”等同于“无房”。将两个数据集进行初步处理后,以“拖欠总得分_离散”为目标,性别、年龄、婚姻状态、户籍、教育程度、职业类别、工作年限、个人收入_连续、保险缴纳、车辆情况、房产为输入,在 SPSS Modeler 工具中建立相应的类型节点。整个分析拖欠程度的数据挖掘流如图 6.56 所示。

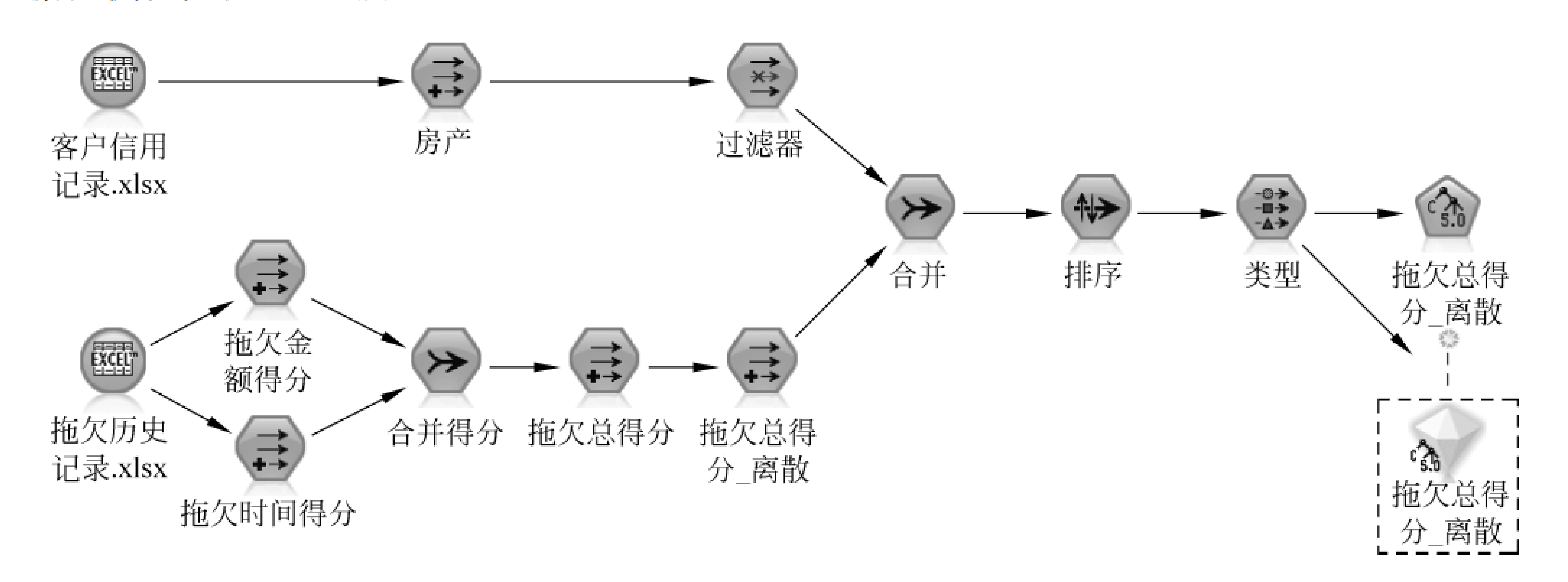


图 6.56 整个分析拖欠程度的数据挖掘流

分析属性重要性发现,户籍和个人收入对拖欠程度影响最大,车辆情况影响较大,性别、年龄和保险缴纳影响较小,其他因素几乎没有影响,其中,个人收入、车辆情况都可以反映一个人的经济实力,因此,拖欠程度和客户的经济实力最相关,其次是户籍、年龄、性别和保险缴纳。

因为“拖欠总金额”与客户的总收入有关,客户的总收入越高,越能申请到高额度的信用



卡,继而容易产生大金额的拖欠,而低收入的客户因为额度的关系,可能很难发生相应金额的拖欠,所以在上述分析拖欠程度时,很显然与客户收入相关,但“逾期天数”和客户总收入并不十分相关,因此可以单独考虑。

在拖欠历史记录数据表中对客户的“逾期天数”进行离散化评估得到“拖欠时间_离散”,如图 6.57 所示。

公式:

```
if (逾期天数 <= 30) then '轻度逾期'
  else if (逾期天数 <= 90) then '中度逾期'
    else if (逾期天数 <=115) then '重度逾期'
      else '严重逾期'
    endif
  endif
endif
```

图 6.57 逾期天数离散化评估

然后对客户信用记录表进行初步处理,将“居住类型”中的“自购房”等同于“有房”“租房”和“其他”等同于“无房”。

将两个数据集进行初步处理后,以“拖欠时间_离散”为目标,性别、年龄、婚姻状态、户籍、教育程度、职业类别、工作年限、个人收入_连续、保险缴纳、车辆情况、房产为输入条件建立相应的类型节点。分析拖欠时间的数据挖掘流如图 6.58 所示。

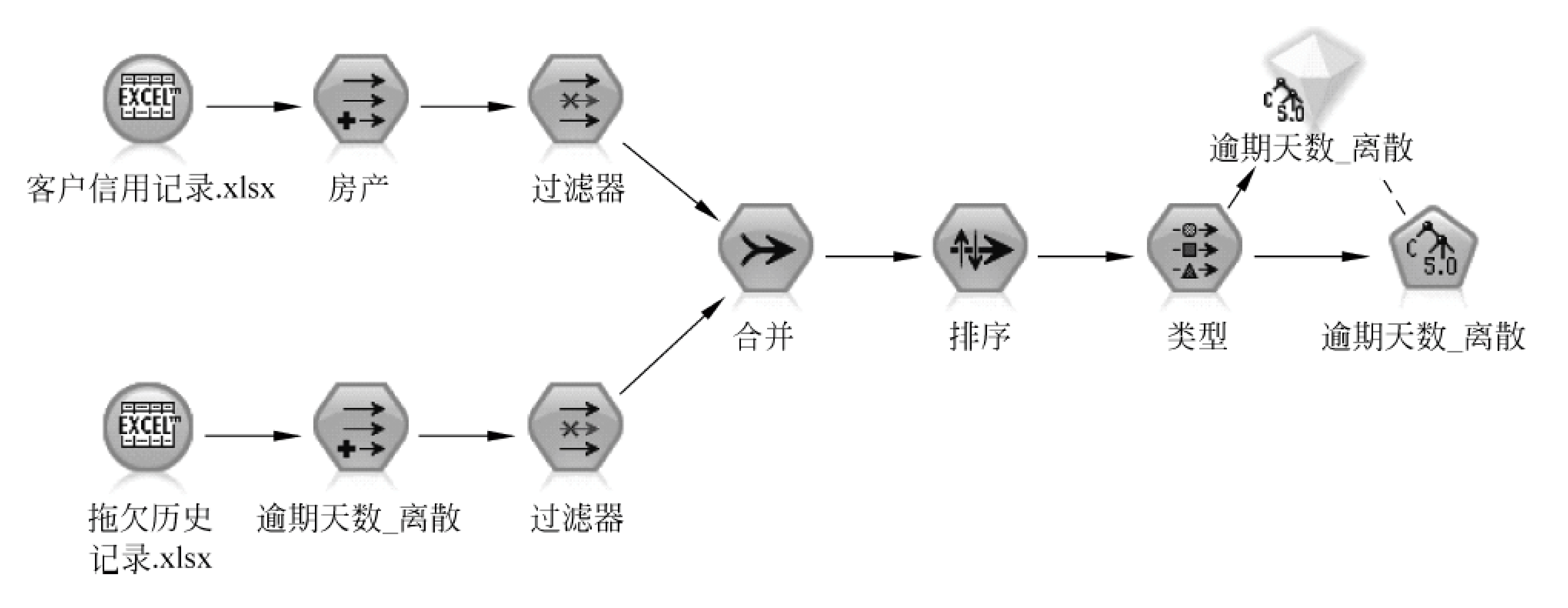


图 6.58 分析拖欠时间的数据挖掘流

在预测变量重要性分析中可以看到工作年限对拖欠时间影响最大,个人收入影响较大,户籍、职业类别、教育程度和性别影响较小。

如图 6.59 所示,在决策树中可以发现具有哪些特征的客户比较容易长时间拖欠。

(1) 工作年限≤4 的客户(约占拖欠客户总数的 30%)大多数为轻度逾期,但对于教育程度为“大专”“高中”和“硕士及以上”的客户有重度逾期的倾向。

(2) 工作年限>4 的客户(约占拖欠客户总数的 70%)大多数为重度逾期,户籍此时对客户的逾期倾向影响很大,呈现明显的地域性差异。

银行对用户进行信用评分时,可以酌情增加工作年限的比重。因为工作年限虽然在欺诈的判断模型中重要性不是很高,但在用户的拖欠时间评判中却有很大的重要性。

```
工作年限 <= 4 [模式: 轻度逾期]
    教育程度 = 初中及以下 [模式: 轻度逾期] ⇨ 轻度逾期
    教育程度 = 大专 [模式: 重度逾期] ⇨ 重度逾期
    教育程度 = 本科 [模式: 轻度逾期]
    教育程度 = 硕士及以上 [模式: 严重逾期]
    教育程度 = 高中 [模式: 严重逾期] ⇨ 严重逾期
工作年限 > 4 [模式: 重度逾期]
    户籍 in ["上海"] [模式: 中度逾期]
    户籍 in ["内蒙古"] [模式: 中度逾期] ⇨ 中度逾期
    户籍 in ["北京"] [模式: 中度逾期]
    户籍 in ["吉林" "新疆" "西藏"] [模式: 严重逾期] ⇨ 严重逾期
    户籍 in ["四川" "天津" "广东" "江苏" "江西" "甘肃" "辽宁"] [模式: 重度逾期] ⇨ 重度逾期
    户籍 in ["宁夏"] [模式: 重度逾期]
    户籍 in ["安徽"] [模式: 重度逾期]
    户籍 in ["山东"] [模式: 重度逾期]
    户籍 in ["山西"] [模式: 重度逾期]
    户籍 in ["广西"] [模式: 重度逾期]
    户籍 in ["河北"] [模式: 重度逾期]
    户籍 in ["河南"] [模式: 重度逾期]
    户籍 in ["浙江"] [模式: 重度逾期]
    户籍 in ["海南"] [模式: 中度逾期]
    户籍 in ["湖北"] [模式: 重度逾期]
    户籍 in ["湖南"] [模式: 重度逾期]
    户籍 in ["福建"] [模式: 重度逾期]
    户籍 in ["贵州"] [模式: 重度逾期]
    户籍 in ["重庆"] [模式: 中度逾期]
    户籍 in ["陕西"] [模式: 重度逾期]
    户籍 in ["青海"] [模式: 轻度逾期]
    户籍 in ["黑龙江"] [模式: 中度逾期]
```

图 6.59　C5.0 决策树分析拖欠时间的决策树

6.5.5　基于回归分析逾期客户特征

通过回归分析用户的人口属性对拖欠行为的具体影响。因为“拖欠总金额”和客户的总收入有关，客户的总收入越高，越能申请到高额度的信用卡，继而容易产生大金额的拖欠，而低收入的客户因为额度的关系，也产生不了较大金额的拖欠，拖欠时间更能够反映用户的拖欠程度，拖欠时间越久，用户越可能不履行还款的义务，银行损失这笔贷款的可能性就越大。这里以用户是否拖欠为因变量来分析逾期客户的特征，见表 6.5。

表 6.5　数据来源与说明

变量类型	变量名	详 细 说 明	取 值 范 围	备　　注
因变量	是否拖欠	0-未拖欠；1-拖欠	0/1	只取整数
自变量：客户因素	性别	定性变量(2 水平)	男、女	男性占比 71.01%
	年龄	单位：岁	18～80	只取整数
	婚姻状况	定性变量(4 水平)	离异/丧偶/未婚/已婚	未婚占比 65.12%
	户籍	定性变量(30 水平)	全国各省	
	教育程度	定性变量(5 水平)	初中及以下/高中/大专/本科/硕士及以上	本科占比 49.36%
	居住类型	定性变量(3 水平)	租房/自购房/其他	租房占比 68.74%
	职业类型	定性变量(5 水平)	个体户/国有企业/私营企业/外资企业/其他企业	私营企业占比 59.71%
	工作年限	单位：年	0～50	只取整数
	保险缴纳	定性变量(2 水平)	有/无	有占比 66.75%
	车辆情况	定性变量(2 水平)	有/无	无占比 65.85%

使用“合并”节点，将“客户号”作为合并的关键字，将信用记录和逾期记录表进行合并，选择“包含匹配和不匹配的记录(完全外部连接)”，并对重复的字段进行滤除，将拖欠历史中的客户号、卡号、额度进行过滤，结果如图 6.60 所示。

从 合并 节点预览 (17 个字段，10 条记录) #1

	拖欠标识	拖欠总...	逾期天数	性别	年龄_连续	婚姻状态	户籍	教育程度	居住
1	1.000	74173....	45.000	男	25.000	未婚	上海	本科	租房
2	$null$	$null$	$null$	男	25.000	未婚	西藏	本科	自购
3	1.000	34055....	68.000	女	25.000	未婚	江西	本科	租房
4	1.000	48222....	99.000	男	32.000	未婚	陕西	本科	租房
5	1.000	75950....	12.000	女	52.000	未婚	河北	本科	租房
6	$null$	$null$	$null$	男	35.000	离异	重庆	硕士及以上	自购
7	$null$	$null$	$null$	男	48.000	已婚	北京	本科	自购
8	$null$	$null$	$null$	男	42.000	未婚	黑龙江	大专	自购
9	$null$	$null$	$null$	男	59.000	已婚	山东	大专	自购
10	1.000	76161....	28.000	男	52.000	已婚	广东	大专	自购

图 6.60 “合并”节点结果预览

可以看到，未发生拖欠的用户记录中，拖欠相关的字段为 null，需要应用“填充”节点对这些字段进行填充，使用“填充”节点，将“拖欠标识”“拖欠总金额”“逾期天数”字段设置为填入字段，替换选项选择“空值”，替换为 0，如图 6.61 所示。



图 6.61 “填充”节点属性设置

为了减少户籍字段取值多对算法的影响，使用“重新分类”节点将各省份聚集为“华北”“华中”“华南”“西北”“西南”“东北”“华东”等大区，如图 6.62 所示。

由于拖欠用户数占比极少，所以在“分区”节点中使用 80%的记录作为训练集，20%的记录作为测试集，增加“自动分类”节点，设置字段的目标变量和输入变量，并使用分区，如图 6.63 所示，在“专家”选项卡中选择所有分类模型。



图 6.62　重新分类节点属性设置



图 6.63　自动分类模型选择

运行自动分类获得效果最好的 3 个分类模型，结果如图 6.64 所示。对比输入变量包含省份或地理区域两种情况下的结果，发现模型的总体精确性几乎没有差别，但对结果模型详情查看，发现输入变量的显著性方面，以省份作为变量效果更佳。



图 6.64 “自动分类”运行结果

选择结果较优的逻辑回归作为客户特征分析模型，详细的分析过程如图 6.65 所示，其中逻辑回归模型中的输入变量和目标变量与“自动分类”模型相同，在“专家”选项卡中选择“专家”模式，并且评估各个预测变量的重要性。

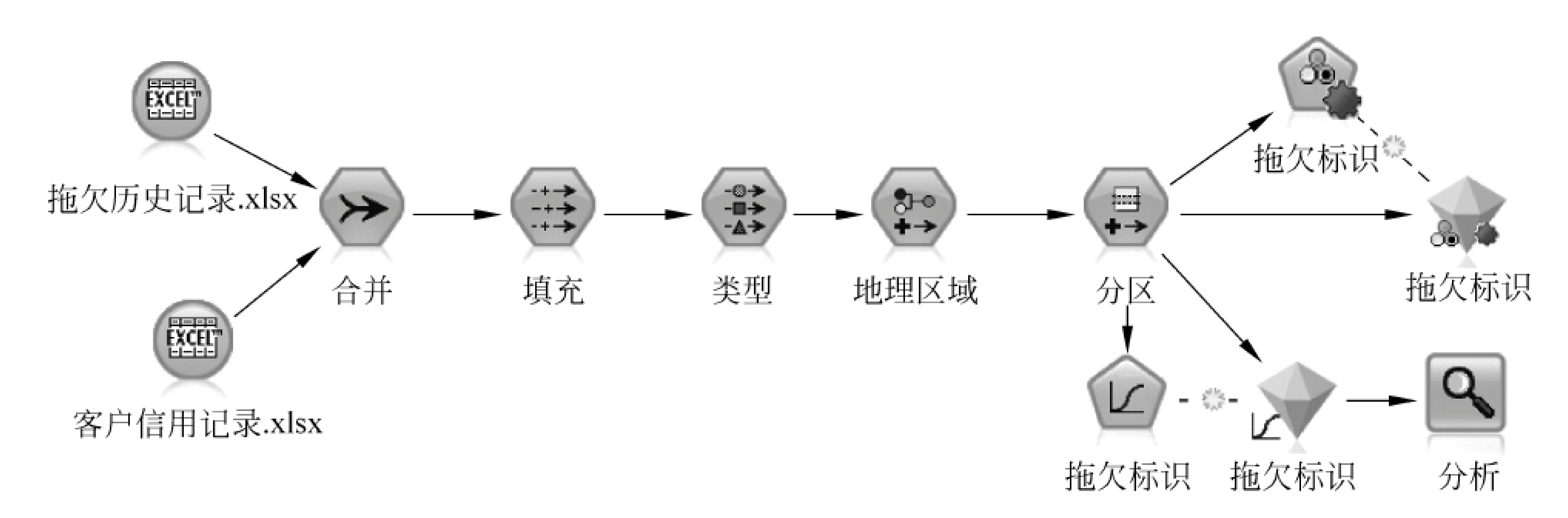


图 6.65 基于逻辑回归分析逾期客户特征

运行逻辑回归模型并查看生成的模型结果，如图 6.66 所示，可以看到其显著性 Sig 指标为 0，但模拟 R 方指标偏低，说明拟合较差。

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1924.811	1931.278	1922.811			
Final	1845.185	2174.968	1743.185	179.626	50	.000

Pseudo R-Square

Cox and Snell	.037
Nagelkerke	.111
McFadden	.093

图 6.66 回归模型拟合性能

图 6.67 显示了回归分析模型的具体结果，图中展示了各个变量的详细影响。从图 6.67 中可以得出具有较高显著性的变量(Sig 指标低于 0.05)为：教育程度(本科、高中)、居住类型(其他)、保险缴纳(无)、信用等级(良好)、户籍(安徽、广东、河北、湖北、湖南、西藏)。无保险缴纳的用户拖欠高于有保险缴纳的用户。

Parameter Estimates

拖欠标识[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
1.0	Intercept	-5.412	.901	36.110	1	.000			
	个人收入_连续	.000	.000	1.843	1	.175	1.000	1.000	1.000
	年龄_连续	-.010	.008	1.633	1	.201	.990	.976	1.005
	[教育程度=本科]	.785	.314	6.236	1	.013	2.193	1.184	4.062
	[教育程度=初中及以下]	.503	.385	1.703	1	.192	1.653	.777	3.518
	[教育程度=大专]	.138	.356	.151	1	.697	1.148	.572	2.305
	[教育程度=高中]	1.208	.363	11.056	1	.001	3.347	1.642	6.823
	[教育程度=硕士及以上]	0[b]	.	.	0	.	.	.	.
	[居住类型=其他]	1.125	.426	6.979	1	.008	3.080	1.337	7.095
	[居住类型=自购房]	.283	.555	.259	1	.611	1.327	.447	3.939
	[居住类型=租房]	0[b]	.	.	0	.	.	.	.
	[职业类别=个体户]	.321	.261	1.513	1	.219	1.379	.826	2.301
	[职业类别=国有企业]	-.654	.436	2.250	1	.134	.520	.221	1.222
	[职业类别=其他企业]	.487	.336	2.104	1	.147	1.627	.843	3.140
	[职业类别=私营企业]	.119	.212	.312	1	.577	1.126	.743	1.707
	[职业类别=外资企业]	0[b]	.	.	0	.	.	.	.
	[保险缴纳=无]	.449	.209	4.592	1	.032	1.566	1.039	2.360
	[保险缴纳=有]	0[b]	.	.	0	.	.	.	.
	[车辆情况=无]	1.012	.519	3.796	1	.051	2.751	.994	7.614
	[车辆情况=有]	0[b]	.	.	0	.	.	.	.
	[信用等级=A-优质客户]	-.535	.473	1.280	1	.258	.586	.232	1.479
	[信用等级=B-良好客户]	-.852	.252	11.417	1	.001	.427	.260	.699
	[信用等级=C-普通客户]	-.075	.148	.259	1	.611	.927	.694	1.240
	[信用等级=D-风险客户]	0[b]	.	.	0	.	.	.	.
	[户籍=安徽]	1.651	.653	6.396	1	.011	5.211	1.450	18.733
	[户籍=北京]	.609	.691	.776	1	.378	1.838	.474	7.119
	[户籍=福建]	1.054	.695	2.301	1	.129	2.868	.735	11.193
	[户籍=甘肃]	.478	.779	.377	1	.539	1.613	.350	7.427
	[户籍=广东]	1.424	.628	5.144	1	.023	4.153	1.213	14.213
	[户籍=广西]	.274	.781	.123	1	.726	1.315	.284	6.079
	[户籍=贵州]	.781	.728	1.148	1	.284	2.183	.524	9.102
	[户籍=海南]	.987	.706	1.956	1	.162	2.683	.673	10.703
	[户籍=河北]	1.789	.640	7.815	1	.005	5.983	1.707	20.971
	[户籍=河南]	.718	.724	.984	1	.321	2.050	.496	8.464
	[户籍=黑龙江]	1.149	.693	2.744	1	.098	3.153	.810	12.273
	[户籍=湖北]	1.534	.661	5.378	1	.020	4.636	1.268	16.950
	[户籍=湖南]	1.539	.662	5.398	1	.020	4.659	1.272	17.064
	[户籍=吉林]	.670	.746	.805	1	.370	1.953	.452	8.432
	[户籍=江苏]	.810	.746	1.176	1	.278	2.247	.520	9.705
	[户籍=江西]	.813	.705	1.327	1	.249	2.254	.565	8.980
	[户籍=辽宁]	.055	.783	.005	1	.944	1.057	.228	4.905
	[户籍=内蒙古]	.328	.779	.178	1	.673	1.389	.302	6.394
	[户籍=宁夏]	1.304	.683	3.642	1	.056	3.682	.965	14.046
	[户籍=青海]	1.223	.675	3.282	1	.070	3.397	.905	12.754
	[户籍=山东]	1.026	.695	2.182	1	.140	2.791	.715	10.895
	[户籍=山西]	.252	.778	.105	1	.746	1.286	.280	5.913
	[户籍=陕西]	1.291	.674	3.665	1	.056	3.637	.970	13.642
	[户籍=上海]	.773	.649	1.419	1	.234	2.165	.607	7.721
	[户籍=四川]	1.036	.682	2.308	1	.129	2.817	.740	10.722
	[户籍=天津]	1.033	.693	2.222	1	.136	2.810	.722	10.926
	[户籍=西藏]	1.839	.875	4.418	1	.036	6.291	1.132	34.955
	[户籍=新疆]	.517	1.179	.192	1	.661	1.677	.166	16.917
	[户籍=浙江]	1.060	.693	2.343	1	.126	2.887	.743	11.224
	[户籍=重庆]	0[b]	.	.	0	.	.	.	.
	[性别=男]	.091	.156	.339	1	.560	1.095	.807	1.485
	[性别=女]	0[b]	.	.	0	.	.	.	.
	[婚姻状态=离异]	.396	.306	1.673	1	.196	1.486	.815	2.707
	[婚姻状态=丧偶]	-17.085	.000	.	1	.	3.802E-8	3.802E-8	3.802E-8
	[婚姻状态=未婚]	.107	.157	.462	1	.497	1.113	.818	1.515
	[婚姻状态=已婚]	0[b]	.	.	0	.	.	.	.

图 6.67 逾期回归分析结果

6.5.6 根据消费历史分析客户特征

信用卡业务能够给银行带来巨大的利益，同时也存在潜在的风险。信用卡业务的两面性要求银行对用户进行分类，按照用户的价值和用户的风险对用户细分，从而对不同类别的用户采取不同的营销措施。

6.5.7 基于聚类分析客户特征

对客户的日常信用卡消费统计数据进行挖掘，从而实现客户分类。对客户细分，可应用聚类分析，详细的过程如图 6.68 所示。首先对数据进行审核，查看数据的完整性和分布特点，然后应用自动聚类选择合适的聚类方法，经过比较发现，K-Means 算法的区分度最高，所以应用这种算法对客户数据进行聚类。

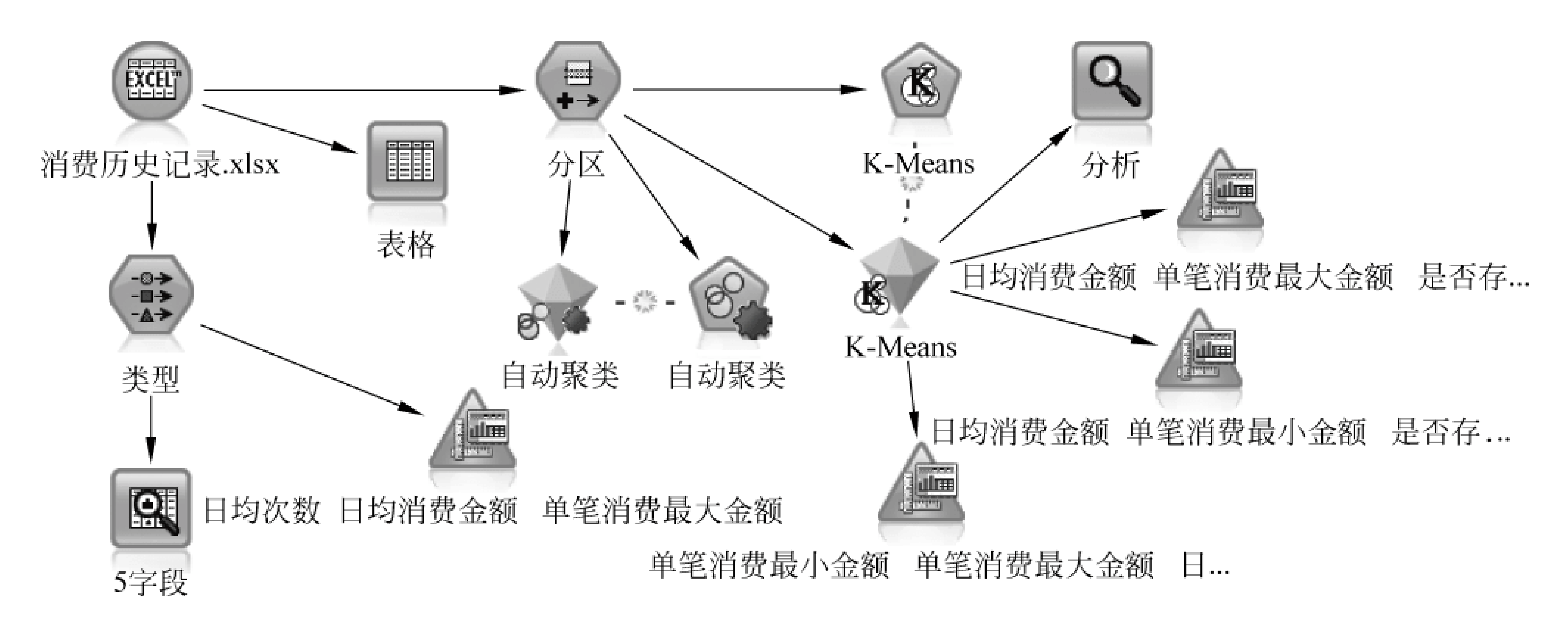


图 6.68 客户聚类分析流程图

为了分析各变量的完整性和数据分布特点，应用数据审核节点对消费历史数据进行探索，如图 6.69 所示。可以看到，数据分布并不符合标准正态分布，特别是个人收入变量相差较大，标准差也很大。

字段	样本图形	测量	最小值	最大值	平均值	标准差	偏度	唯一	有效
日均消费金额		连续	30.000	81797.000	8588.844	19410.212	2.851	--	5954
日均次数		连续	1.000	28.000	3.017	2.139	3.256	--	5954
单笔消费最小金额		连续	1.500	6052.300	767.326	1520.346	2.390	--	5954
单笔消费最大金额		连续	30.300	500000.000	32985.127	61030.249	5.663	--	5954
个人收入_连续		连续	10416.000	99000000000.000	58940691.965	2226598764.237	44.241	--	5954

图 6.69 变量质量审核结果

单击“质量”界面，查看各输入变量的质量，如图 6.70 所示，没有数值缺失、空值、空白值等问题。

完整字段(%): 100% 完整记录(%): 100%

字段	测量	离群值	极值	操作	缺失插补	方法	完成百分比	有效记录	空值	字符型空值	空白	空白值
日均消费金额	连续	0	545	无	从不	固定	100	5954	0	0	0	0
日均次数	连续	201	37	无	从不	固定	100	5954	0	0	0	0
单笔消费最小金额	连续	24	690	无	从不	固定	100	5954	0	0	0	0
单笔消费最大金额	连续	126	122	无	从不	固定	100	5954	0	0	0	0
个人收入_连续	连续	19	951	无	从不	固定	100	5954	0	0	0	0

图 6.70 各变量的质量审核结果

在图形板中应用散点图矩阵对日均消费金额、日均次数、单笔消费最小金额、单笔消费最大金额进行可视化显示，并以是否存在欺诈行为作为标记，较大的圆形表示有欺诈行为，如图 6.71 所示。

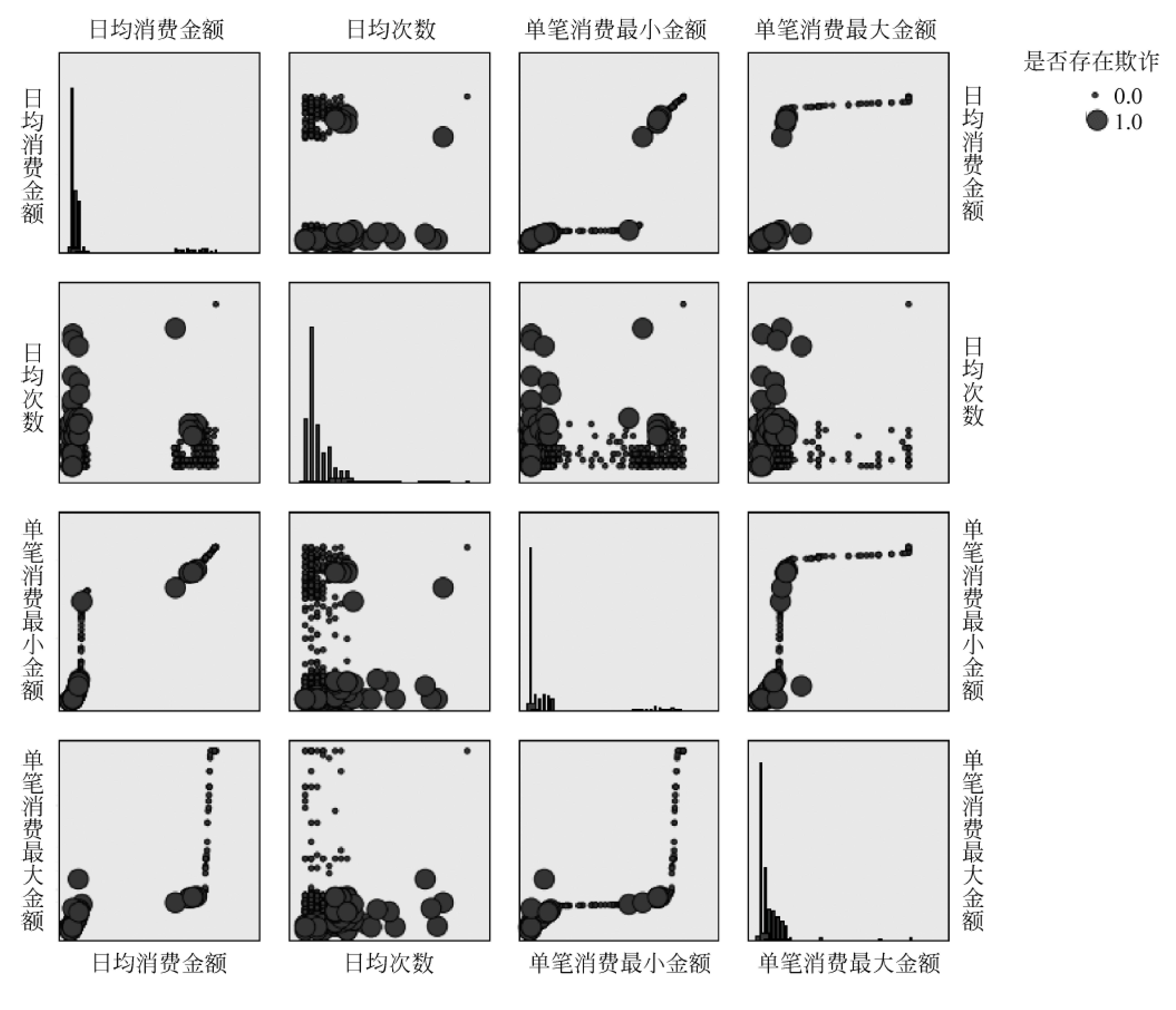


图 6.71 各变量的散点图矩阵

4 个变量形成矩阵关系，图 6.71 中的 16 个图形左下角与右上角为横、纵坐标对称。通过观察日均次数与日均消费金额散点图，可以看到呈现明显的聚类效应，且日均消费金额较低的客户欺诈行为较多。

单笔消费最小金额与日均消费的散点图中，两个聚类均具有一定的线性关系，随着日均消费金额的增长，单笔消费最小金额也在快速提高，但基增长率在下降，即随着日均消费能力不断提高，单笔消费最小金额增长变化较慢。

单笔消费最大金额与日均消费金额散点图中，随着日均消费金额的增长，单笔消费最大金额呈现先慢后快的趋势。

日均次数与单笔消费最小金额、单笔消费最大金额的散点图中，日均次数不具有区分能力，但单笔消费最小金额具有更高的区分度，其中单笔消费最大金额和单笔消费最小金额较低的情况下欺诈行为较多。

单笔消费最小金额和单笔消费最大金额的散点图中，两者在不同阶段呈现不同的线性关系，在单笔最低消费金额增长的情况下，单笔消费最大金额变化并不明显，但达到奇点时，单笔消费最大金额呈几何级增长。

从各散点图中簇类分布情况，可以看出日均消费金额、单笔消费最小金额、单笔消费最大金额均具有较强的分类能力，如图 6.72 所示。

自动聚类

文件(F) 生成(G) 预览(P)

模型 摘要 注解

排序方式(S): 使用 升序 降序 删除未使用模型 视图: 测试集

是否使...	图形	模型	构建时间(分钟)	轮廓	聚类数	最小聚类(N)	最小聚类(%)	最大聚类(N)	最大聚类(%)	最小/最大	重要性
☑		K-m...	< 1	0.897	5	6	0	2624	86	0.002	0.0
☐		两步 1	< 1	0.820	2	461	15	2571	84	0.179	0.0
☐		Koh...	< 1	0.461	11	6	0	1031	34	0.006	0.0

确定 取消 应用(A) 重置(R)

图 6.72　各变量的散点图矩阵

在 K-Means 聚类中选择日均消费金额、日均次数、单笔消费最小金额和单笔消费最大金额作为输入字段，聚类数选 5 个，在“专家”选项卡中选择专家模式，参数为默认值，如图 6.73 所示。



图 6.73　K-Means 聚类自变量选择

运行模型得到聚类的结果，可以看到模型的聚类质量较高，达到 0.8 的轮廓系数值，如图 6.74 所示。

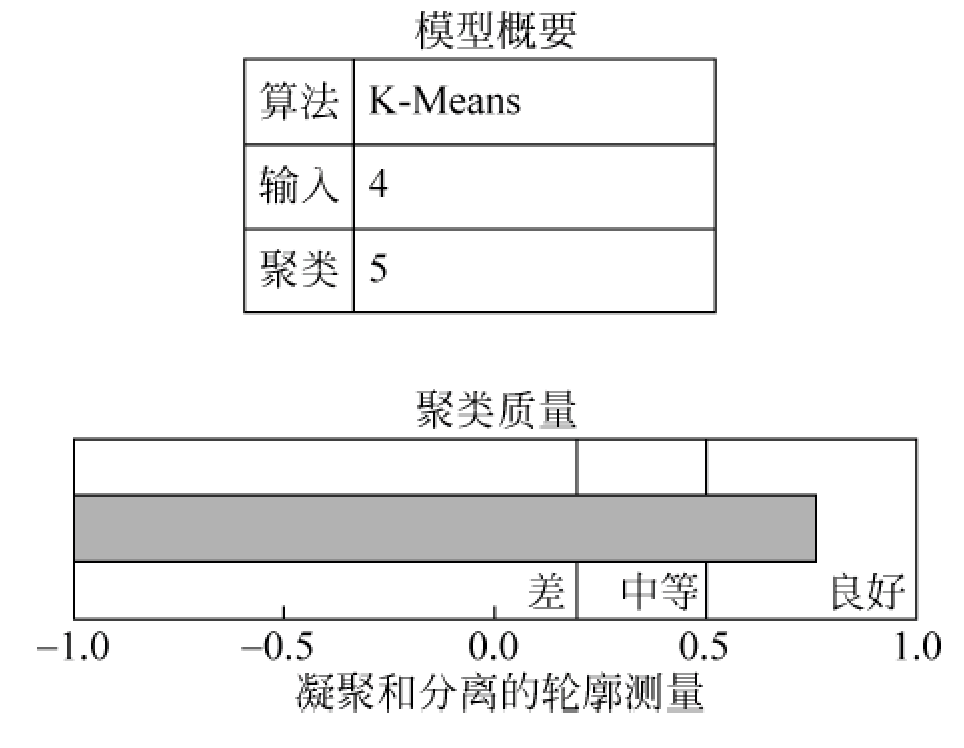


图 6.74　K-Means 聚类模型结果

查看聚类的大小，发现最大的类别占比为 88.3%，最小的类别占比只有 0.1%，说明聚类的类别数量并不合理。通过观察聚类中各变量在聚类中的重要性和区分度，发现日均消费金额、单笔消费最大金额和单笔消费最小金额的重要性基本一致，除 88.3%之外的几个簇的区别并不明显，如图 6.75 所示。

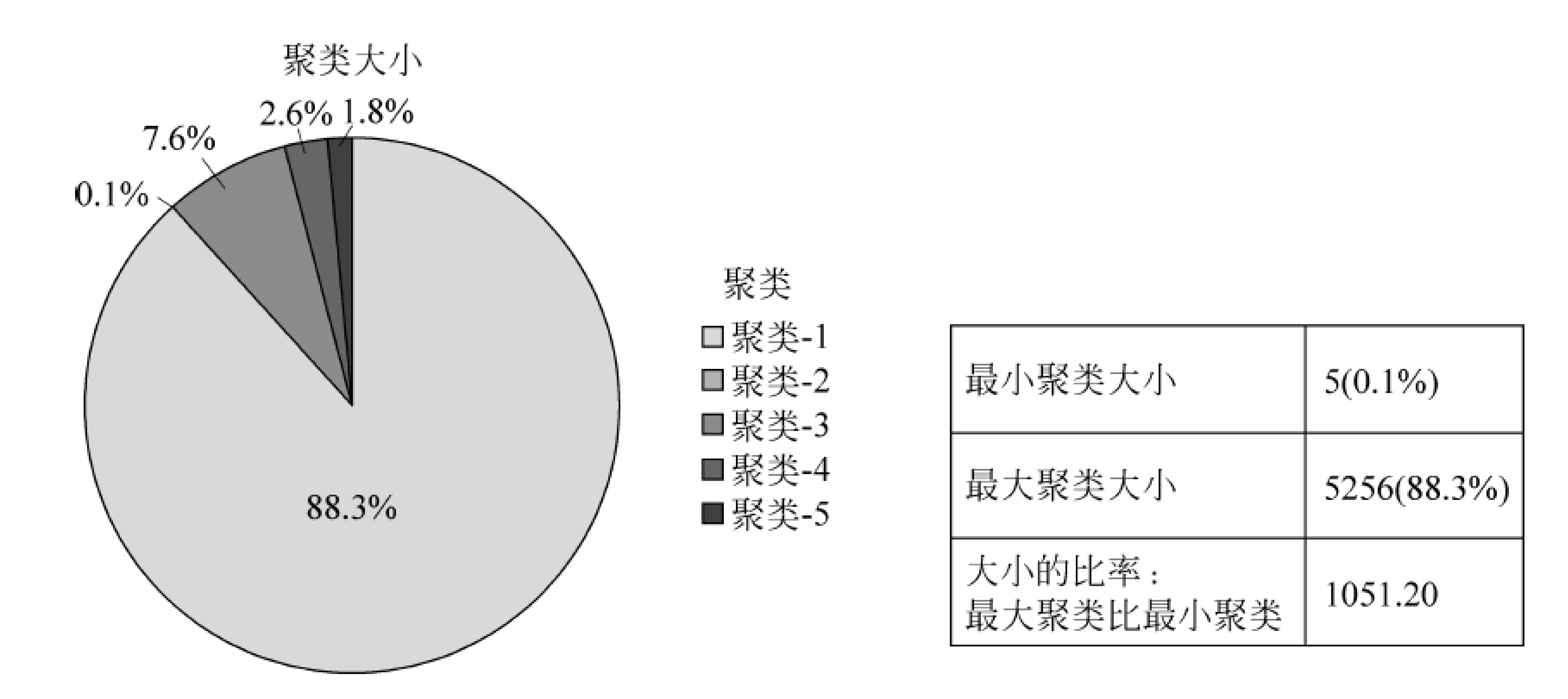


图 6.75　K-Means 聚类各类别分布情况

保留模型的其他参数不变，K-Means 的聚类数量改为 2，并重新运行模型，得到新的模型结果，达到 0.9 的轮廓系数。图 6.76 是两种聚类的详细分类依据，从中可以看出 90%的用户单笔消费最大金额低于 2 万元，单笔消费最小金额低于 326 元，日均消费金额少于 2488 元，可以视为一般客户，除此之外的 10%用户为优质客户。

为了查看两个簇类下各自变量的分布情况，同时选中两个类别，在聚类比较界面可以看到不同类在单笔消费最大、单笔消费最小、日均消费金额中均有较大的不同，而日均次数几乎没有差别，这说明日均次数重要程度最低，如图 6.77 所示。

为了查看聚类中最重要的两个自变量之间的关系，使用图形板分析两者的散点图分布，用点的大小区分是否存在欺诈行为，大的圆点表示存在欺诈行为。如图 6.78 所示，用户分为两类：日均消费低于 1 万元，单笔消费最小金额低于 4500 元为一个簇类 A，而日均消费高于 60 000 元，单笔最小金额高于 4500 元为另外一个簇类 B。其中，类 A 中单笔消费最小

输入(预测变量)重要性

■1.0 ■0.8 ■0.6 ■0.4 ■0.2 □0.0

聚类	聚类-1	聚类-2
标签		
描述		
大小	90.8% (5409)	9.2% (545)
输入	单笔消费最大金额 20 473.52	单笔消费最小金额 5145.18
	单笔消费最小金额 326.22	日均消费金额 69 134.74
	日均消费金额 2488.36	单笔消费最大金额 157 159.97
	日均次数 3.00	日均次数 3.17

图 6.76 输入变量的分类阈值

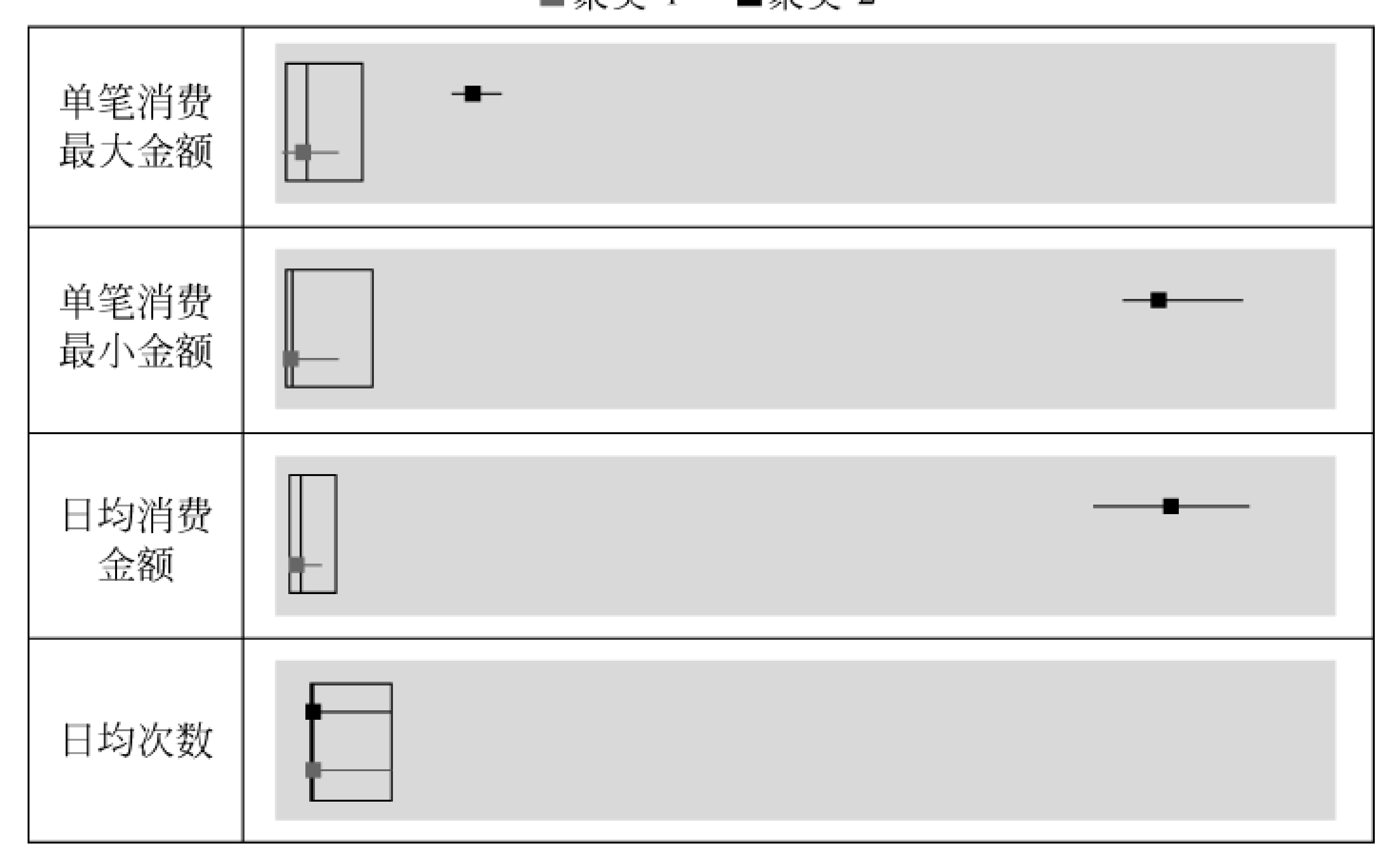


图 6.77 不同聚类的输入变量分布情况

金额低于 1000 元的用户中,存在更多的欺诈行为,需要重点关注。而类 B 中单笔消费最小金额高于 5000 元的用户无欺诈行为,说明这部分人群为优质客户中的最优客户。

如图 6.79 所示,可将用户分为以下两类用户:日均消费低于 10 000 元,单笔消费最小金额低于 4400 元,单笔消费最大金额低于 9000 元;日均消费高于 59 000 元,单笔消费最小金额高于 4400 元,单笔消费最大金额高于 9000 元。其中,后者为优质客户,前者为一般客户。

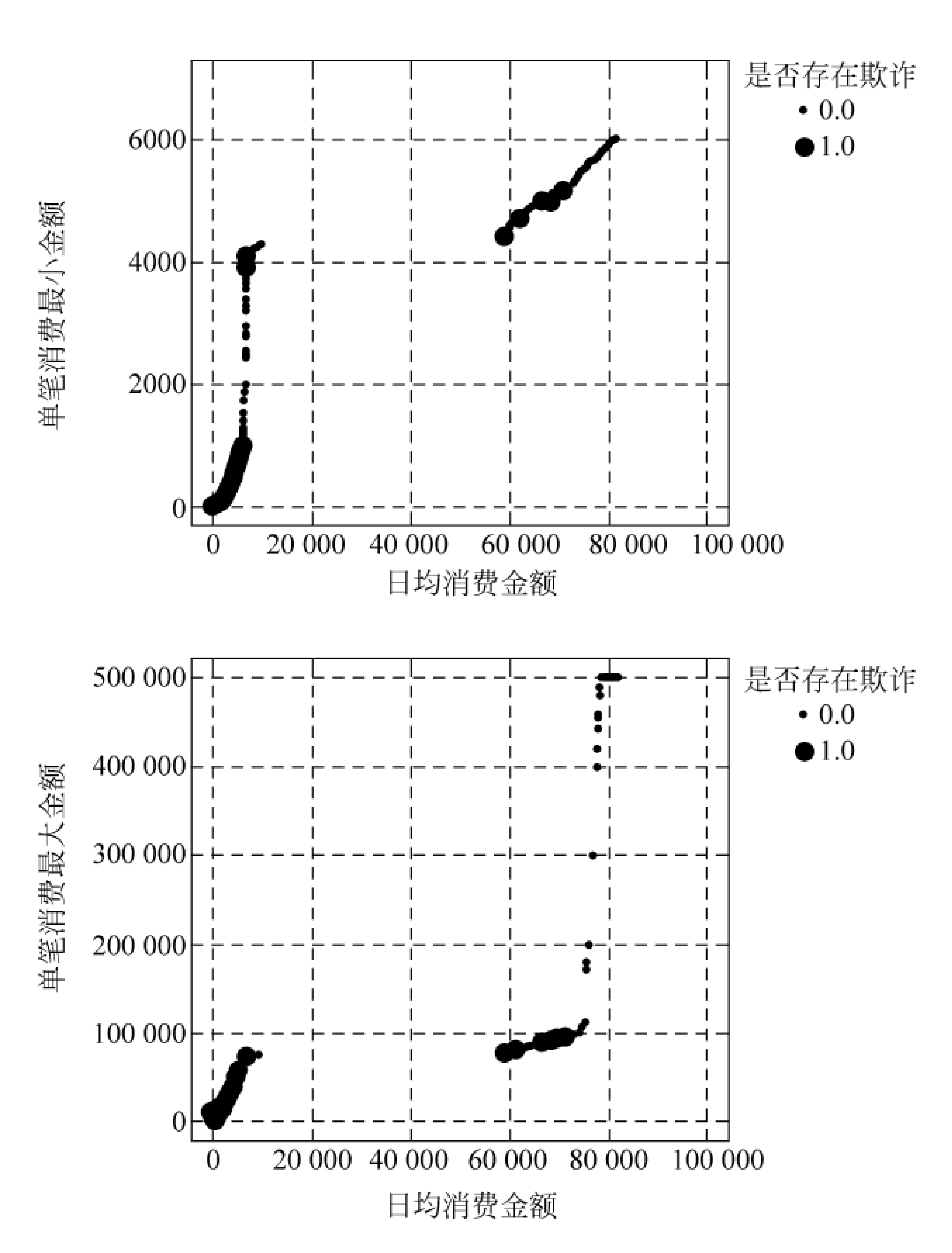


图 6.78　日均消费金额、单笔消费最大金额散点图

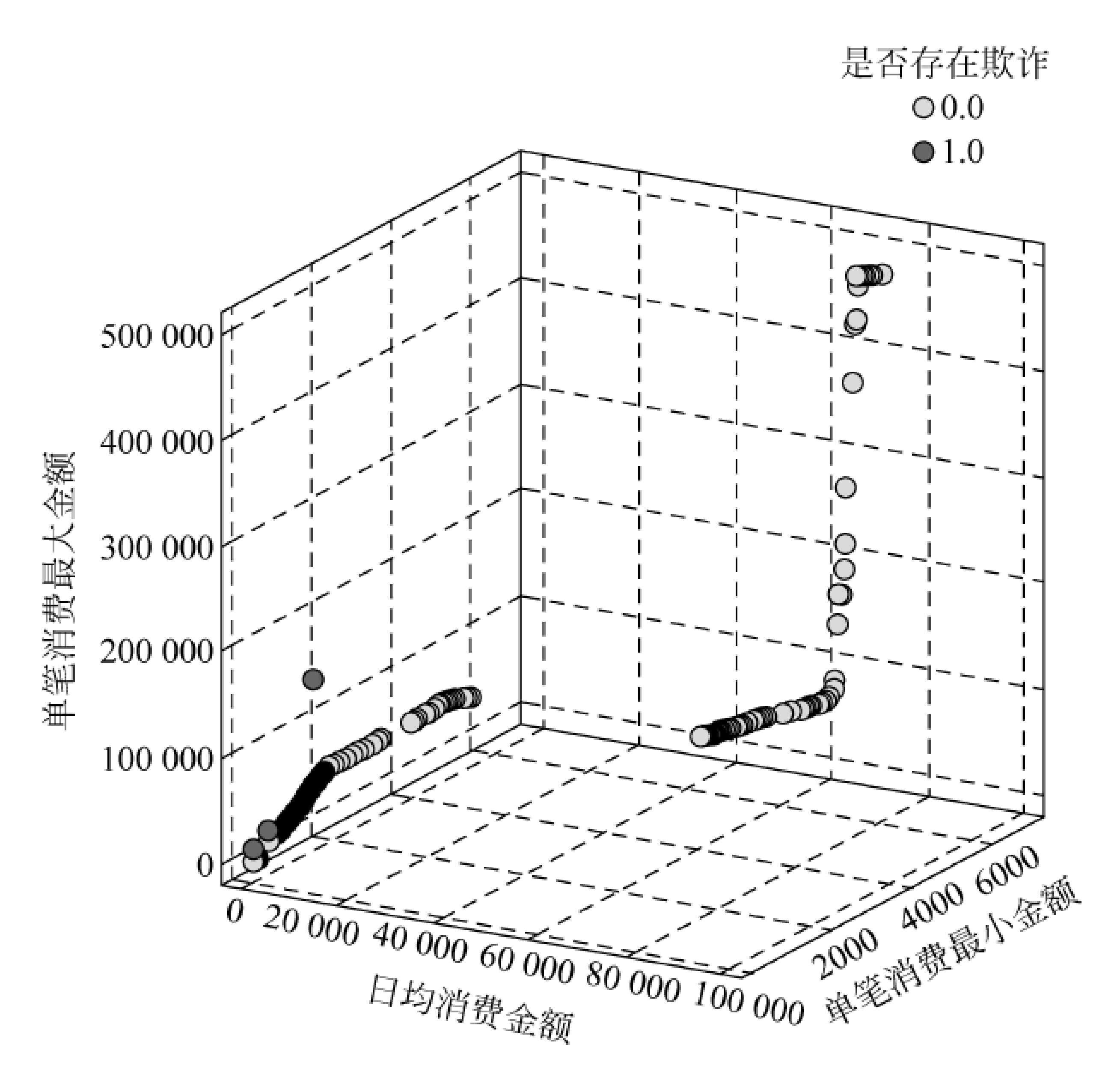


图 6.79　日均消费金额、单笔消费最大金额、单笔消费最小金额的三维散点图

使用三维散点图展示日均消费金额、单笔消费最大金额、单笔消费最小金额之间的关系，以及由此构成的聚类特征，可以直观地看到 3 个变量在两种类别人群中的分布情况。

6.5.8 基于客户细分的聚类分析

根据用户的历史消费记录，通过用户的日均消费金额、日均次数等可以划分用户给银行带来的价值，通过用户是否存在欺诈、拖欠得分(由拖欠金额和拖欠时间综合得到)、信用评分可以衡量用户存在和潜在的风险。信用评分是对用户潜在风险的一个总体体现，包含了用户的人口属性。对每一个持卡人可以划分 5 类特征，分别是日均消费金额、日均次数、是否欺诈、拖欠得分和信用评分。

持卡人的 5 类特征可以分别进行排序。其中，日均消费金额、日均次数、信用评分 3 个特征将数据分级为 5 部分，并对每一部分的客户赋予 1～5 的值。例如，对于日均消费金额，最高的一组用户赋值为 5，中间的 3 组用户分别赋予 4、3、2，日均消费金额最低的一组用户值为 1。这样处理之后，记日均消费金额得分为 M，日均次数得分为 F，信用评分得分为 C。对于这三类特征的用户，得分越高说明客户的价值越高，或者风险越低。

是否欺诈和拖欠得分两类特征的计算中，将未产生欺诈和拖欠的用户特征值记为 0。为了加重欺诈行为对用户的惩罚，将产生欺诈的用户得分记为 5，无欺诈用户设置为 0。而拖欠的用户按照拖欠得分的高低，从低到高分别赋值为 1～5。记处理后的欺诈特征值为 A，拖欠得分特征值为 D。对于这两类特征，得分越高说明用户的风险越高。M/F/C/A/D 特征计算方法如图 6.80 所示。

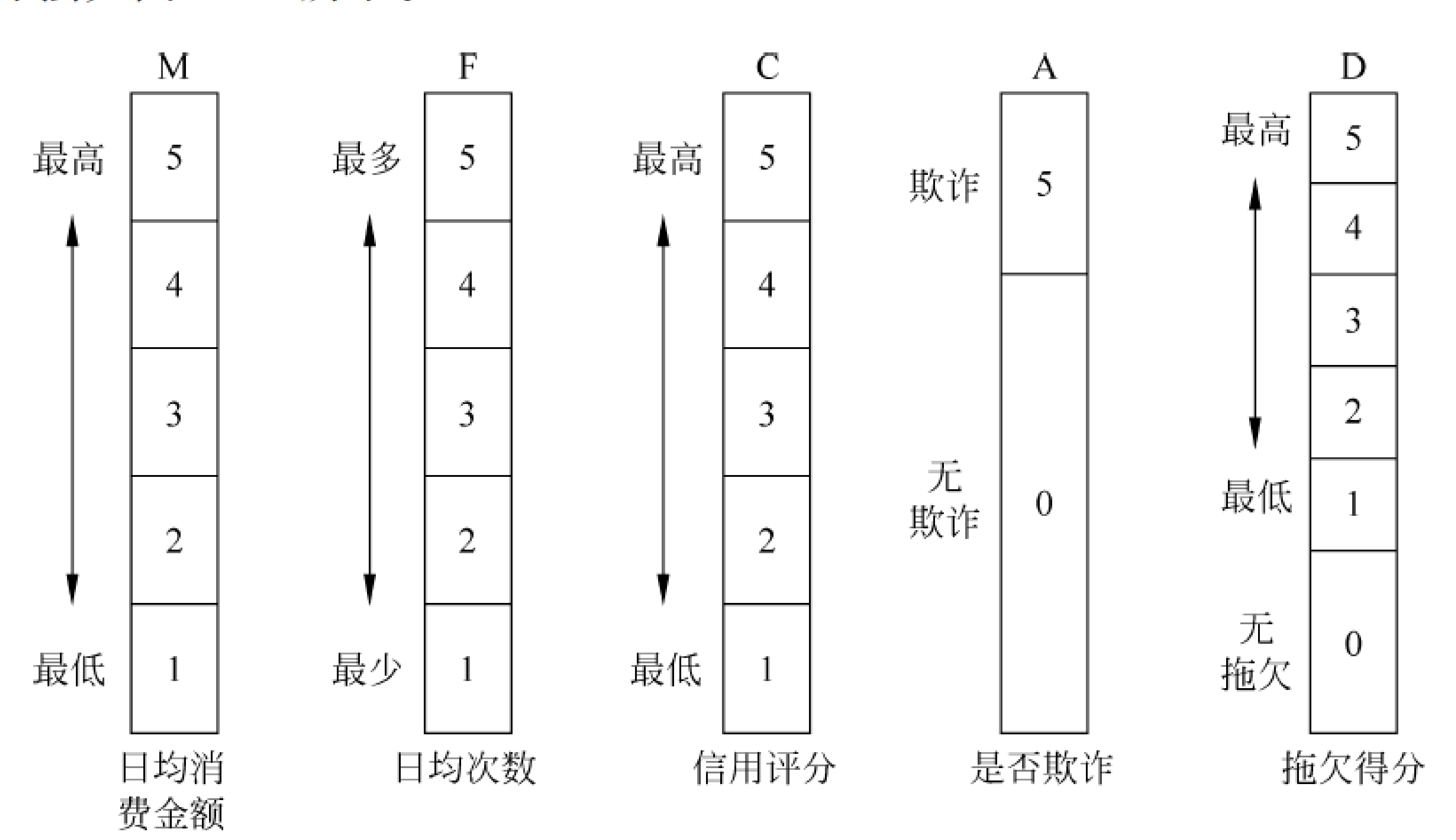


图 6.80 M/F/C/A/D 特征计算方法

使用“过滤器”节点，将“客户号”“卡号”等标识用户个人的变量过滤。删除本次分析的无效变量“拖欠标识”“拖欠总金额”等字段，只剩下与用户人口属性有关的字段，如图 6.81 所示。

使用“分级”节点，将用户的日均消费金额、日均次数、信用评分、是否欺诈、拖欠得分进行分级处理，得出具体的 M、F、C、A、D 等级值，每级拥有相同的用户数量，使用 K-Means 算法按照用户的价值和风险分析用户的具体分类。“分级”节点属性设置如图 6.82 所示。



图 6.81　“过滤器”节点属性设置



图 6.82　“分级”节点属性设置

基于这种特征计算方法，M、F 值均高的为高价值客户，均低的为低价值客户；C 值高，A、D 值均低的为低风险客户；C 值低，A、D 值均高的为高风险客户。根据获取的用户交易数据，计算每个用户的 M、F、C、A、D 值，调用 K-Means 聚类算法将用户聚为 9 个簇，如图 6.83 所示。

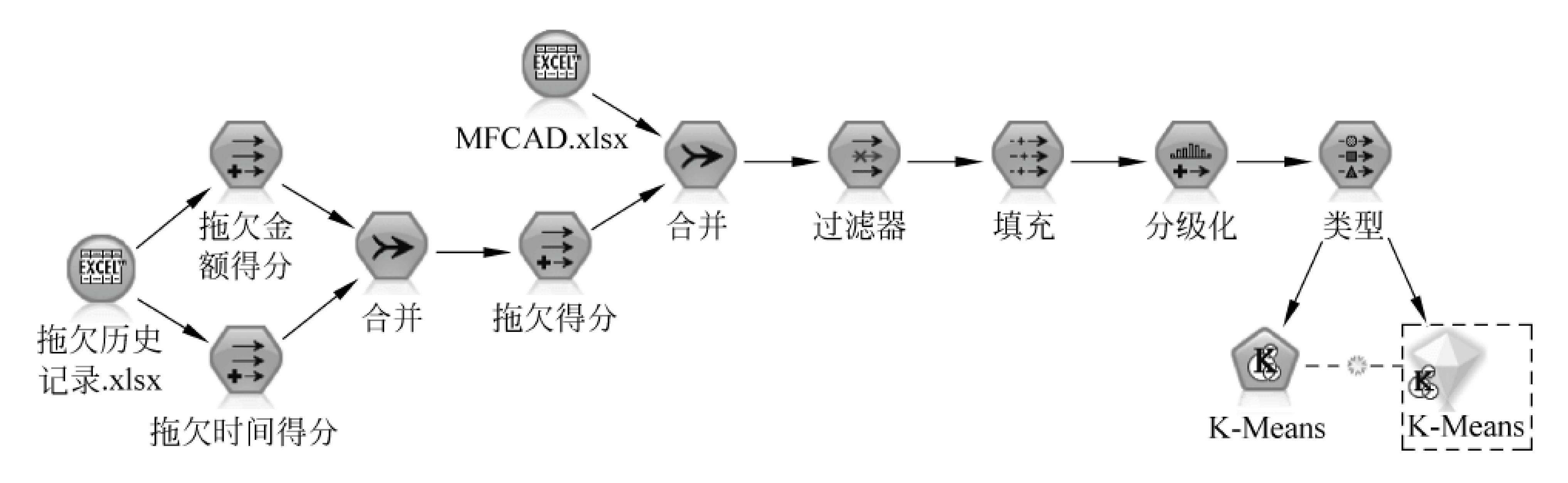


图 6.83　用户分类流程图

设置聚类数量为 9，图 6.84 显示模型的聚类效果良好，在可以接受的范围内。表 6.6 显示了聚类后结果得到 9 个簇，各个簇的编号以及对应的 5 个特征值。

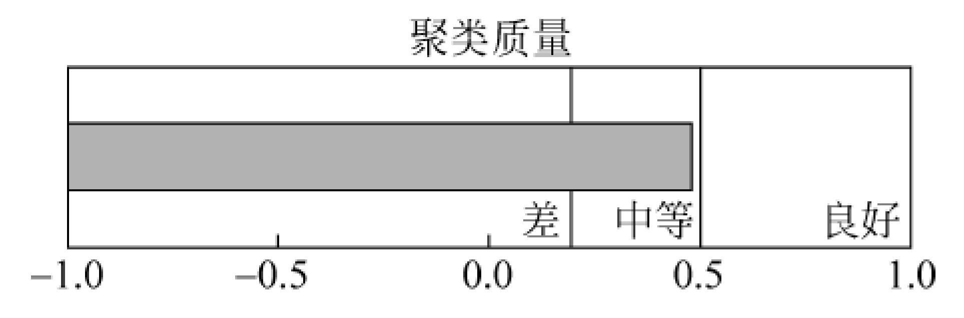


图 6.84 聚类模型质量

表 6.6 M/F/C/A/D 特征值聚类结果

簇号	M	F	C	A	D
1	1.81	2.00	1.75	5.00	0.01
2	4.41	4.99	3.90	5.00	4.02
3	4.15	4.66	4.08	0.00	0.01
4	3.15	4.94	1.49	5.00	1.85
5	2.48	3.04	3.08	0.00	3.64
6	1.42	3.92	2.69	5.00	3.12
7	4.01	1.99	4.04	0.00	0.00
8	2.48	3.04	3.08	0.00	3.64
9	1.75	4.68	1.95	0.00	0.01

对于得到的 9 个簇，每个簇对应一类用户，用户按照价值和风险分类，如图 6.85 所示分为 9 类。

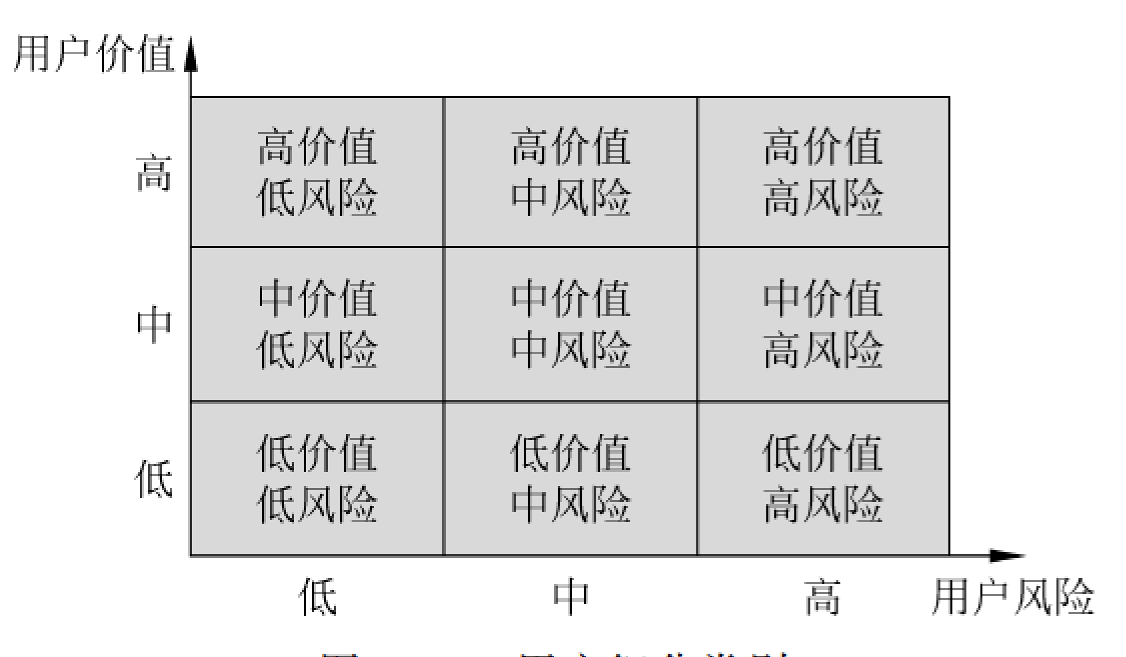


图 6.85 用户细分类别

按照 5 个特征值的定义方法，M、F 值越高的用户，价值越高，反之价值越低；C 值越高的用户，风险越低，反之风险越高；A、D 值越高的用户，风险越高，反之风险越低。通过各个聚类的特征值比较，将 9 个簇分别划分类别，结果见表 6.7。

表 6.7 用户分类及对应聚类簇号

价值/风险	高风险	中风险	低风险
高价值	2	4	3
中价值	6	8	7
低价值	1	5	9

对得到的9类用户，分别统计各类用户在总用户中的人数和所占百分比，得到的结果见表6.8。

表6.8　各类别用户数量及所占比例

簇号	用户类别	用户数量/人	用户占比/%
1	低价值、高风险	1961	32.9
2	高价值、高风险	100	1.7
3	高价值、低风险	699	11.8
4	高价值、中风险	111	1.9
5	低价值、中风险	25	0.4
6	中价值、高风险	26	0.4
7	中价值、低风险	2298	38.6
8	中价值、中风险	31	0.5
9	低价值、低风险	703	11.8

这样的分类考虑了用户的价值和风险，引入用户的人口属性来评估风险，可以有效地考虑到用户潜在的风险性，为银行的信用卡业务管理提供了参考。对不同类别的用户制定不同的营销策略，可达到良好的服务效果。例如，应当为“高价值、低风险”的用户提供优质的服务，挽留这样的用户，以防其流失。对于“低价值、高风险”的用户，则应当适当加强管控，提高服务费。

第 7 章

海底捞火锅运营分析

随着社会的不断发展，人们的生活水平不断提高，去餐馆吃饭已经从过去的奢侈享受变成了现在的家常便饭，各种新的餐馆、饭店也如雨后春笋般不断涌现，饮食行业竞争愈发激烈，并且越来越呈现出白热化的趋势。

自 20 世纪 80 年代中期起，火锅企业开拓创新发展，尤其是近几年来，火锅业的迅猛发展引起全社会的关注。其中，火锅老字号企业焕发新春，再塑辉煌。新型火锅企业锐意进取，异军突起。火锅企业的连锁经营步伐逐渐加快，连锁店网点数量不断增加，连锁经营的区域也日益拓展，企业规模和实力不断增强，知名品牌不断涌现。

行业的快速发展也带来许多问题，火锅菜品加工工艺相对简单，非常容易复制。市场上只要出现一款畅销的菜品，很快各个店都竞相模仿，导致目前火锅行业菜品单一化现象严重，没有在原料和工艺上对菜品进行创新。由于进入火锅行业的门槛较低，对从业人员的要求并不高，随之而来的是从业人员整体素质相对落后，没有过硬的专业技术，服务理念、经营管理理念、复合管理能力欠佳，从而影响了整个行业的服务水平。大量的新店不断涌现，其中不乏盲目跟风者，导致惨淡经营，给火锅业造成负担，同时也使得火锅店之间的竞争日趋激烈。

在企业的众多经营活动中，每天都会产生大量的数据，这些看似毫无关联的数据，往往具有深层次的紧密关系，对企业的经营和发展策略的决策都会有十分重要的作用和意义。随着大数据时代的来临，数据分析已经成为企业的经营管理者极重视的一项活动。数据分析可以对客观情况进行正确的反映，对企业经营管理过程中产生的数据进行监督，能够有效地改善企业进行各项活动的决策。本章以海底捞火锅店（北京北太平庄牡丹园店）为例进行数据分析。分析饭店的相关数据，同时与同行竞争对手做比较，为饭店的未来发展以及营销提出建议。

7.1 火锅相关数据抓取

利用 Python 脚本作为数据抓取工具。BeautifulSoup 库(可以通过 pip 下载)提供找到 HTML 中标签的方法,利用标签得到标签下的文本信息或者标签的属性信息,抓取海底捞(牡丹园店)的数据,使用脚本为 again. py。为了在抓取的过程中使脚本更像是人为的操作,而不是爬虫在工作,需要设置好请求头(Request Header)中的参数。

这里设置了很多备用参数,在使用过程中随机换备用参数,可以适当地提高在 IP 被封禁之前抓取的数据量。在“大众点评”的一条用户评论中可以根据分析的需要,抓取多项数据,如图 7.1 所示。这些数据包括用户昵称,用户的贡献值,用户对这次用餐的总评分(平均评分),用户对这次用餐的口味、环境、服务的评价,用户的评论内容,用户的用餐时间(评论时间),用户这条评论收到的点赞数等。



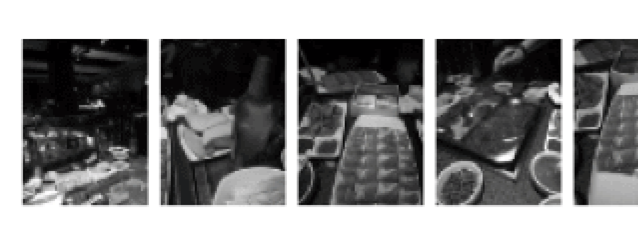


图 7.1 抓取页面内容

根据这些数据在 HTML 页面中的标签信息编写代码,利用 find、find_all 方法找标签,其中第一个参数是标签的名称,第二个参数是标签的属性值,find 方法是找到符合筛选条件的第一个标签,而 find_all 方法是找到符合筛选条件的所有标签的一个数组。寻找用户昵称标签,然后将标签内的文本内容添加到事先定义好的 name 数组中。star 中存储的是用户的贡献值。用户的贡献值在网页中以标签属性的形式存在,通过 span 的 class 名来反映。time、score、environment、serve、taste、comment、zan 依次是存储时间、总分、环境、服务、口味、评论、点赞数的数组。

把抓取的数据存储到数组后,利用 Python 读写 Excel 的库将数据存入 Excel 表。将 Excel 的对应表格的值设置为对应的已抓取数据。

接着抓取用户喜欢的菜的数据,这个数据也位于用户评论的页面中,将这项数据与其他数据分开抓取,是因为喜欢的菜这项内容属于选填项,而且有很大一部分用户不去填写喜欢的菜,如果与其他的数据一同抓取,会导致出现较多的数据空白。采集的页面如图 7.2 所示,采集数据有昵称、时间、喜欢的菜等。

查看页面源代码,发现需要抓取数据的标签及其属性。在脚本运行的过程中不仅仅是要抓取一个页面中的内容,而是要抓取很多结构与标签相似的页面内容,因此写一个循环,自动访问页面的下一页,一种方法是根据页面中的标签文本内容得到下一页面的 URL,也可以在写代码时将 URL 直接输入。



咩小咩_

人均 ¥100 | 口味4(非常好) 环境4(非常好) 服务4(非常好)

【位置】在牡丹园地铁站出来步行10分钟左右，位置挺好找的。周边有一个潮汕粥吧好像，停车场车位挺多的。
【环境】店面很大，很干净整洁。门口等位区有五子棋折星星，还有很多小吃可以吃。而且小吃可以打包带走~还有一个小二层，是儿童玩耍区域~
【服务】服务一直是海底捞的主打啊，去厕所的路上遇到服务员会主动打招呼;厕所有专门的阿姨帮你递擦手纸;叫服务员时候随叫随到，会主动帮你添加水啊，下锅什么的。门口的小吃可以打包带回家吃哦~
【口味】首先说，餐具都变小了...原来盘子是平的，现在盘子变成这样了~小料儿碗也变浅了...南瓜粥是无糖的，喜欢甜的可以叫服务员给白糖;菌汤超好喝;小料儿自调的，超级超级好吃！西瓜超甜！！！辣锅超爽！！！总之，真的好好吃！！！！

喜欢的菜：抻面 鸭肠 鲜虾滑 油豆皮 鸳鸯锅 自助小料 内蒙羔羊肉 等位时的各种小吃 巴沙鱼片



查看更多图片»

04-19 海底捞火锅 赞 | 回应 | 收藏 | 不当内容

图 7.2 采集的页面

7.2 数据预处理

将脚本抓取的数据在 Excel 中打开，如图 7.3 所示。

	A	B	C	D	E	F	G	H	I
1	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank4	03-31 更新于17-04-12 12:57	joyhao1985		(3)
2	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank2	04-11	特别喜欢吃川菜		0
3	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank2	04-11	luckywjh		0
4	irr-star5	口味3(很好	服务4(非	环境3(很好	urr-rank5	04-11	dpuser_23928410		0
5	irr-star5	口味4(非	服务2(好)	环境2(好)	urr-rank4	04-10	童殿的舌头		0
6	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank3	04-10	爱吃榴莲的z小姐		0
7	irr-star5	口味3(很好	服务4(非	环境4(非	urr-rank2	04-09	pngfng		0
8	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank2	04-09	时光煮雨颜依旧		0
9	irr-star4	口味4(非	服务3(很好	环境3(很好	urr-rank5	04-09	dpuser_306081072		0
10	irr-star5	口味4(非	服务4(非	环境3(很好	urr-rank1	04-08	学院路路草		0
11	irr-star5	口味4(非	服务4(非	环境3(很好	urr-rank5	04-08	南悟LSY		0
12	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank5	04-07	小洁_7689		0
13	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank5	04-07	dpuser_82659776		(1)
14	irr-star4	口味4(非	服务4(非	环境4(非	urr-rank4	04-07	cuijiajia		0
15	irr-star5	口味4(非	服务4(非	环境4(非	urr-rank5	04-06	Xa_6623		(1)
16	irr-star4	口味3(很好	服务3(很好	环境3(很好	urr-rank3	04-06	雷少666		0

图 7.3 脚本抓取的数据

可见，原始数据很粗糙，缺少列名称，同时冗余数据和残缺数据都较多，因此进行下列处理。

首先添加列名称：在第一行对数据添加类别说明，分别为"评价均分""口味评分""服务评分""环境评分""用户贡献值""评价日期""用户昵称""评价内容""评价点赞数"；去重处理：全选，单击"数据"-"删除重复项"。得到的数据如图 7.4 所示。

删除无用数据，评论与评分的分析仅需要与评分有关的前 4 列和"评价内容"，所以保留上述列，删除其余列。复制一份删除后的表格文件，对复制后的文件，删除评分相关的 4 列，然后将文件另存为 comments.txt 文件，该 txt 文件即为所有的评论内容。

将数据导入 SPSS Modeler 18.0，单击"插入"→"源"→Excel，选择文件类型和导入文件。因为这里不需要"用户昵称""评论内容"和"评价点赞数"等字段，所以单击"过滤"，过滤这 3 个字段。最后单击"确定"按钮，如图 7.5 所示。

可通过单击"插入"→"输出"→"表"，并运行该表查看导入的数据，如图 7.6 所示。

	A	B	C	D	E	F	G	H	I
1	评价均分	口味评分	服务评分	环境评分	用户贡献值	评价日期	用户昵称	评价内容	评价点赞数
2	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank4	03-31	joyhao1985		(3)
3	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank2	04-11	特别喜欢吃川菜		0
4	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank2	04-11	luckywjh		0
5	irr-star5	口味3(很好	服务4(非常	环境3(很好	urr-rank5	04-11	dpuser_23928410		0
6	irr-star5	口味4(非常	服务2(好)	环境2(好)	urr-rank4	04-10	童殿的舌头		0
7	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank3	04-10	爱吃榴莲的z小姐		0
8	irr-star5	口味3(很好	服务4(非常	环境4(非常	urr-rank2	04-09	pngfng		0
9	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank2	04-09	时光煮雨颜依旧		0
10	irr-star4	口味4(非常	服务3(很好	环境3(很好	urr-rank5	04-09	dpuser_306081072		0
11	irr-star5	口味4(非常	服务4(非常	环境3(很好	urr-rank1	04-08	学院路路草		0
12	irr-star5	口味4(非常	服务4(非常	环境3(很好	urr-rank5	04-08	南悟LSY		0
13	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank5	04-07	小洁_7689		0
14	irr-star5	口味4(非常	服务4(非常	环境4(非常	urr-rank5	04-07	dpuser_82659776		(1)
15	irr-star4	口味4(非常	服务4(非常	环境4(非常	urr-rank4	04-07	cuijiajia		0

图 7.4　Excel 初步预处理后的数据



图 7.5　设置 Excel“源”



表（6 个字段，5,160 条记录）#1

文件　编辑　生成

表　注解

	评价均分	口味评分	服务评分	环境评分	用户贡献...	评价日期
1	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank45	03-31 更...
2	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank20	04-11
3	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank20	04-11
4	irr-star50	口味3(很好)	服务4(非常好)	环境3(很好)	urr-rank5	04-11
5	irr-star50	口味4(非常好)	服务2(好)	环境2(好)	urr-rank40	04-10
6	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank30	04-10
7	irr-star50	口味3(很好)	服务4(非常好)	环境4(非常好)	urr-rank20	04-09
8	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank20	04-09
9	irr-star40	口味4(非常好)	服务3(很好)	环境3(很好)	urr-rank5	04-09
10	irr-star50	口味4(非常好)	服务4(非常好)	环境3(很好)	urr-rank10	04-08
11	irr-star50	口味4(非常好)	服务4(非常好)	环境3(很好)	urr-rank5	04-08
12	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank5	04-07
13	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank5	04-07
14	irr-star40	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank40	04-07
15	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank5	04-06
16	irr-star40	口味3(很好)	服务3(很好)	环境3(很好)	urr-rank30	04-06
17	irr-star50	口味4(非常好)	服务4(非常好)	环境3(很好)	urr-rank20	04-05
18	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank30	04-05
19	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank5	04-05
20	irr-star50	口味4(非常好)	服务4(非常好)	环境4(非常好)	urr-rank40	04-05

确定

图 7.6　使用 SPSS 导入后的数据

从图 7.6 中可以看出,“评价均分”的格式为“irr-star * 0”,其中“ * ”为实际的数值,所以需要将这个数值提取出来。而“口味评分”“服务评分”和“环境评分”的数据虽然不是数值型,但是数据含义明显且已经离散成 5 类,所以不必进行处理。“用户贡献值”的数据与“评价均分”类似,需要将最后的数值提取出来。而对于“评价日期”,元素数据是离散到每一天,这样离散程度太高了。考虑到一般商家是按照月度进行考核,可以将其处理成以月份为单位。

单击“插入”→“字段选项”→“导出”,将“导出字段”设为“评价月份”,需要将原始数据中的评价日期处理为以月份为单位,这里舍弃了部分数据格式不符合规范的数据,并将其处理成−1。在“公式”栏中输入“if(length(评价日期)==5) then substring_between(1,2,评价日期) elseif(length(评价日期)==8) then substring_between(4,5,评价日期) else "−1" endif”,单击“确定”按钮,如图 7.7 所示。



图 7.7 数值化“评价月份”

同样,可以添加“表”来查看字段。将“评价均分”和“用户贡献值”改为数字形式。分别添加“导出”节点,设置“导出字段”和“公式”为“评价均分(数字)”“substring_between(9,9,评价均分)”;“用户贡献值(数字)”“allbutfirst(8,用户贡献值)”。

继续添加“过滤”节点,已经不需要“评价均分”“用户贡献值”和“评价日期”3 个字段了,将其过滤。添加“表”节点,查看现在的数据,如图 7.8 所示。

对菜品内容做预处理,用到的文件为前面抓取到的菜品内容。“爬虫”抓取的原始数据如图 7.9 所示。

将 C 列包含菜名的数据复制并粘贴到新的 Excel 表格中,选择导出为 txt 文本文件,重

口味评分	服务评分	环境评分	评价月份	评价均分（数字）	用户贡献值（数字）
口味4(非常好)	服务4(非常好)	环境3(很好)	04	5	10
口味4(非常好)	服务4(非常好)	环境3(很好)	04	5	5
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	5	5
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	5	5
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	4	40
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	5	5
口味3(很好)	服务3(很好)	环境3(很好)	04	4	30
口味4(非常好)	服务4(非常好)	环境3(很好)	04	5	20
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	5	30
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	5	5
口味4(非常好)	服务4(非常好)	环境4(非常好)	04	5	40
口味4(非常好)	服务4(非常好)	环境3(很好)	04	5	30

图 7.8 处理后的评分数据

	A	B	C	D	E	F	G	H	I	J	K
1	16-12-17	草原上的超	滑牛肉 鸭血 柠檬水 豆浆 金针菇 鲜鸭血 手切羊肉								
2	16-06-29	爱吃饭的F	滑牛肉 一根面 海底捞笋片 午餐肉 鸭肠 虾滑 鱼片 捞面								
3	16-12-16	LvZvTo	鸭血 虾滑 嫩牛肉 鱼片								
4	16-12-12	时光溅落	鸭肠 小料 鲜毛肚								
5	16-11-11	小糊涂_姿	一根面 滑牛肉								
6	16-10-09	Linda_832	滑牛肉 海底捞牛肉								
7	16-10-07	kiyoface	滑牛肉 虾滑 毛肚 油豆皮 鱼丸 小料 猪脑花								
8	16-10-06	一颗荔枝内	蟹棒 鲜虾滑 滑牛肉 无刺巴沙鱼鱼片								
9	16-10-06	dpuser_28	海底捞笋片 虾滑 毛肚 海带 鲜毛肚 豌豆尖 自助小料 龙利鱼片 海底捞血旺 鸭舌头 巴沙鱼								
10	16-09-11	挪若岩2	豆浆 豆皮 手切羊肉 鲜虾滑 竹笋 鸳鸯锅 青笋 西式牛滑 山药 黑豆浆 茼蒿 鸭肠 毛肚 羊肉								
11	16-09-10	嗜血如兰	牛肉 羊肉 一根面 嫩牛肉 虾滑 猪脑								
12	16-09-07	一帆杰作	滑牛肉 一根面 海底捞牛肉 海底捞笋片 嫩牛肉 虾滑								
13	16-09-07	努力的向日	滑牛肉 一根面 海底捞牛肉 鸭血 毛肚								
14	16-09-04	Jay小颖	一根面 嫩牛肉 虾滑								
15	16-08-11	飞扬的hea	牛肉丸 抻面 精品牛肉 撒尿牛肉丸（弹性超好）								

图 7.9 “爬虫”抓取的原始数据

命名为 meal.txt。将该 txt 文件中所有的空白部分替换为单个空格。下面用 Python 脚本处理该 txt，代码如下所示。

```
#!/usr/bin/python
# -*- coding:utf-8 -*-
word_lst = []
word_dict = {}
with open('./meal.txt') as wf,open("word.txt",'w') as wf2:

    for word in wf:
        word_lst.append(word.split(' '))
        for item in word_lst:
            for item2 in item:
                if item2 not in word_dict:
                    word_dict[item2] = 1
                else:
                    word_dict[item2] += 1
    final_dict = sorted(word_dict.items(),key=lambda item: item[1],reverse=True)
    for x,y in final_dict:
        wf2.write(str(x) + " " + str(y) + "\n")
```

上段代码读取 meal.txt 中的菜品名，并统计每种菜品出现的次数。最终输出为 word.txt，如图 7.10 所示，展示了推荐数比较多的一些菜品。

这里的数据预处理利用到前面的 word.txt 和“菜品.xlsx”，用 Python 实现。选取推荐数大于 10 的菜进行关联分析，将每一个菜名设置为新的表格列名称。

```
滑牛肉 2250
海底捞牛肉 1613
一根面 1592
鸭血 1314
柠檬水 1099
抻面 1080
海底捞笋片 807
午餐肉 703
豆花 701
鸭肠 675
虾滑 570
嫩牛肉 225
毛肚 171
鱼片 164
竹荪 152
猪蹄 93
猪脑 90
豆浆 88
牛肉丸 79
捞面 68
羊肉丸 41
拉面 41
羊肉 20
肥牛 19
香蕉酥 18
金针菇 15
简阳鱼 15
手切羊肉 15
番茄锅底 14
小料 14
黄喉 13
凉茶 13
简州鱼 12
```

图 7.10 推荐数较多的菜品

```
import xlwt
import xlrd
#要输出的表格
workbook = xlwt.Workbook()
sheet1 = workbook.add_sheet('sheet1',cell_overwrite_ok = true)
#读取统计的词频
f = open('./word.txt','r',encoding = 'UTF-8')
content = f.readlines()
f.close()
#添加第一行的菜品名
num = 0
writeNum = 0
while num < len(content)
    tem = content[num].find(' ')
    mealCount = content[num][tem: -1]              #商品数量
    if int(mealCount)> = 10:                        #选择数量大于 10 的菜品
        content[num] = content[num][0:tem]
        sheet1.write(0,writeNum,content[num])
        mealTup = mealTup + (content[num],)
        writeNum += 1
    num += 1
```

读取抓取的“菜品.xlsx”中的每一个用户的推荐菜，若列名称中的菜出现在该用户的推荐菜中，则将对应的单元格设为 1，否则设为 0。

```
#读取抓取的数据
workbook1 = xlrd.open_workbook('./菜品.xlsx')
worksheets = workbook1.sheet_names()
worksheet1 = workbook1.sheet_by_name(u'其余的评价')
num_rows = worksheet1.nrows
for curr_row in range(num_rows):
    row = worksheet1.row_values(curr_row) #每一行
    mealFlag = 0
    while mealFlag < len(mealTup):
```

```
            mealName = mealTup[mealFlag]
            try:
                row.index(mealName)
                sheet1.write(curr_row + 1,mealFlag,1)
            except:
                sheet1.write(curr_row + 1,mealFlag,0)
            mealFlag += 1
    workbook.save('meal.xls')
```

处理后的 meal.xls 如图 7.11 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	滑牛肉	海底捞牛肉	一根面	鸭血	柠檬水	抻面	海底捞笋片	午餐肉	豆花	鸭肠	虾滑	嫩牛肉	毛肚
2	1	0	0	1	1	0	0	0	0	0	0	0	0
3	1	0	1	0	0	0	1	1	0	1	1	0	0
4	0	0	0	1	0	0	0	0	0	0	1	1	0
5	0	0	0	0	0	0	0	0	0	1	0	0	0
6	1	0	1	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	1	0	1
9	1	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	0	0	1	0	1
11	0	0	0	0	0	0	0	0	0	1	0	0	1
12	0	0	1	0	0	0	0	0	0	0	1	1	0
13	1	1	1	0	0	0	1	0	0	0	1	1	0
14	1	1	1	1	0	0	0	0	0	0	0	0	1
15	0	0	1	0	0	0	0	0	0	0	1	1	0
16	0	0	0	0	0	1	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	1	1	0	0	0	0	0	0	0	0	0
19	1	0	0	1	0	0	0	0	0	0	0	1	0
20	0	0	0	0	0	0	0	0	0	0	1	0	0
21	0	1	0	0	1	0	0	0	0	1	1	0	1
22	1	1	1	1	0	0	0	0	0	1	1	0	0

图 7.11　处理后的推荐菜品统计

7.3　数据分析

7.3.1　海底捞运营分析

为了能够深入地了解海底捞店铺的经营情况，分析影响该店生意的关键因素，并且通过和其他店铺对比得出海底捞的优势和劣势。首先将处理过的有关海底捞（牡丹园店）的数据导入 Watson，提出问题“What drives 总分?”，结果如图 7.12 所示。

与预期的一致，口味、环境、服务是 3 个主要的影响因素。首先分析口味这一最主要因素的影响。在 Watson 中输入“口味 and 总分”可以看到一些可以提问的问题的提示，选择问题“How does the number of Rows compare by 口味 and 总分?”，得到图 7.13 所示的结果。

图 7.13 反映出在口味为 4 的评价中，绝大部分的顾客都给出了 5 分或者 4 分的总评分数，所以口味对于一家火锅店而言是至关重要的。此时引入时间维度，首先考虑时间维度与口味的关系。“What are the number of each 口味 and 时间?”，如图 7.14 所示。

从图 7.14 看到，明显的事实是 2014 年的顾客要明显多于 2015 年与 2016 年的顾客，这里反映出了这家火锅店存在的问题（接下来会加以分析，此处继续分析口味与时间关系），可以为开设分店以及制定新的策略提供参考。首先通过统计 2013 年以及 2012 年的顾客人数发现 2014 年的人数并不是突然的井喷，而是延续着 2012 年以及 2013 年的销量，所以分析

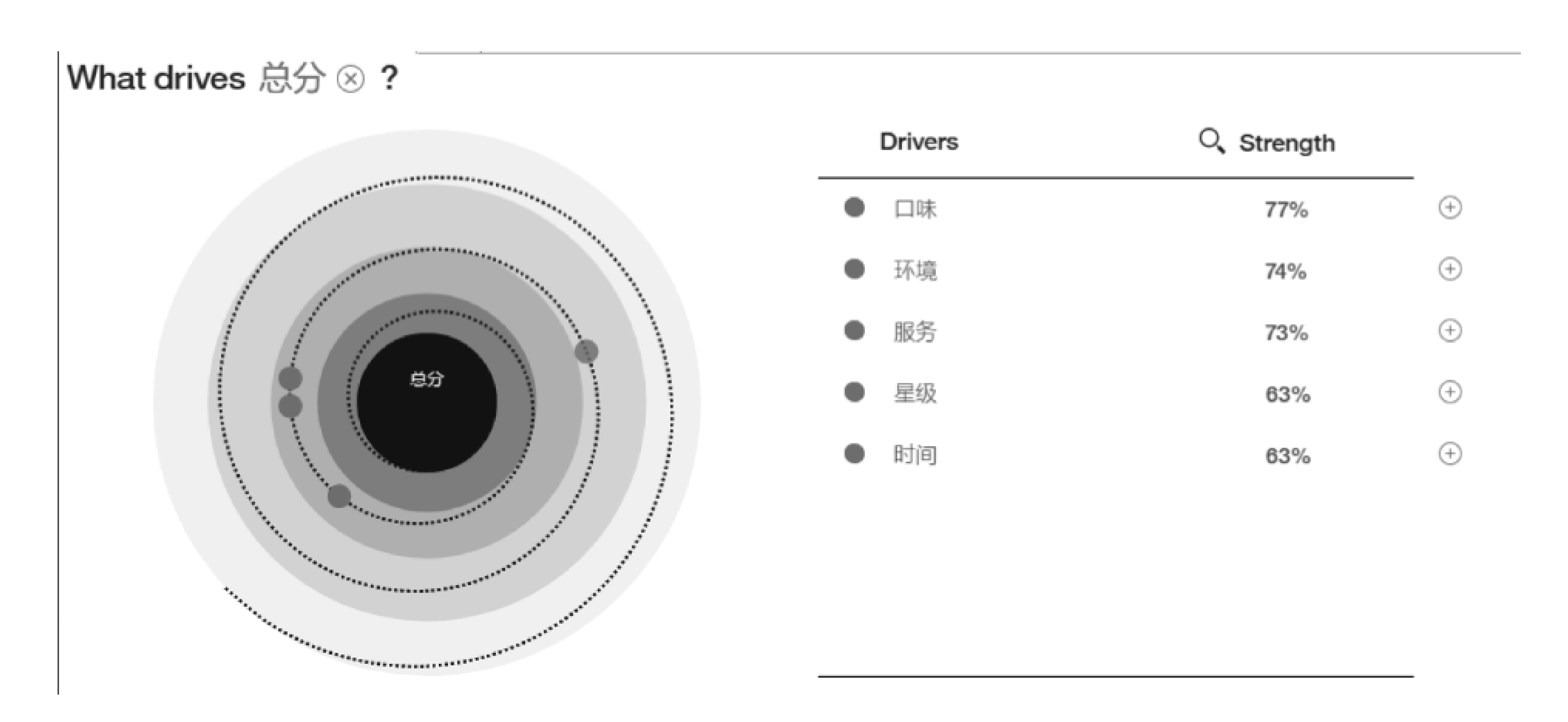


图 7.12　影响海底捞生意的因素

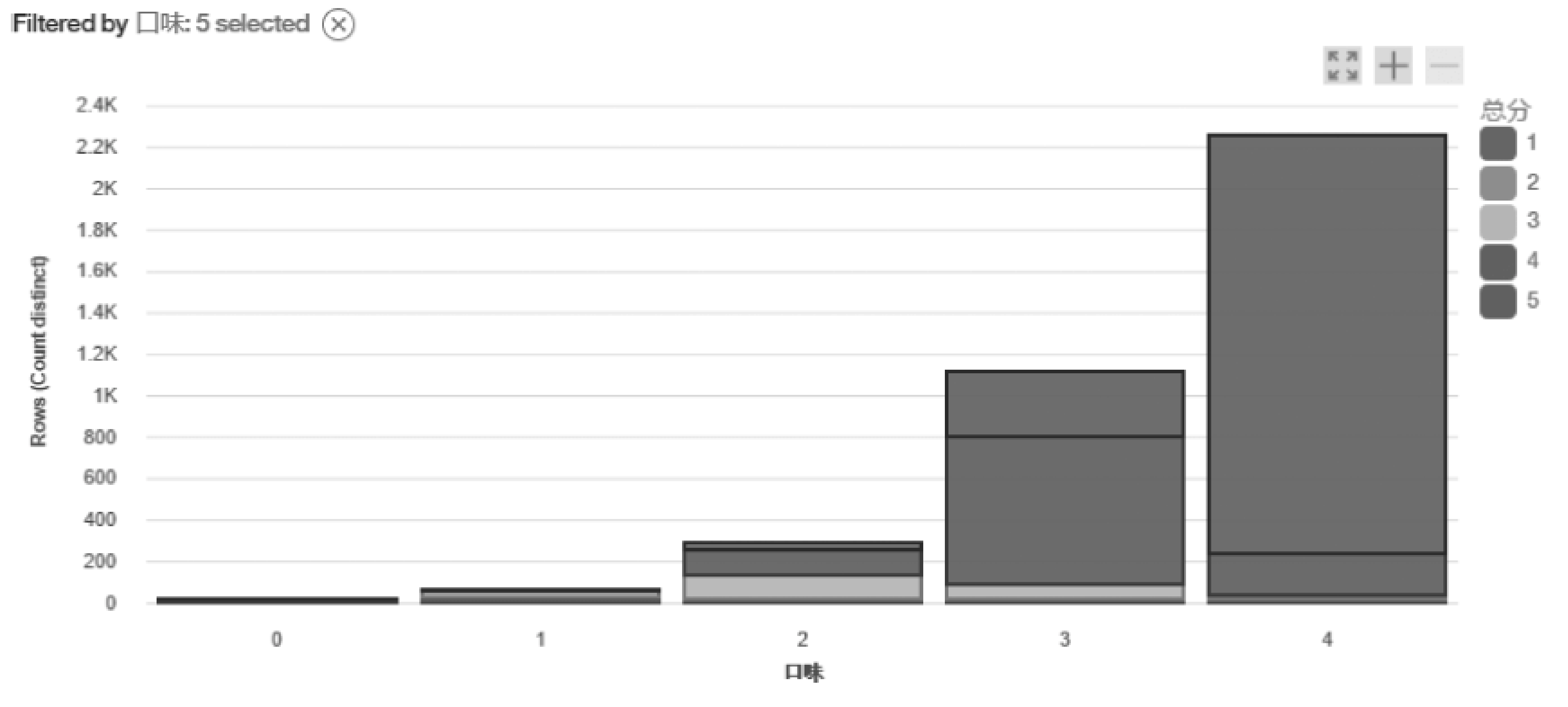


图 7.13　口味以及总分对应数量图

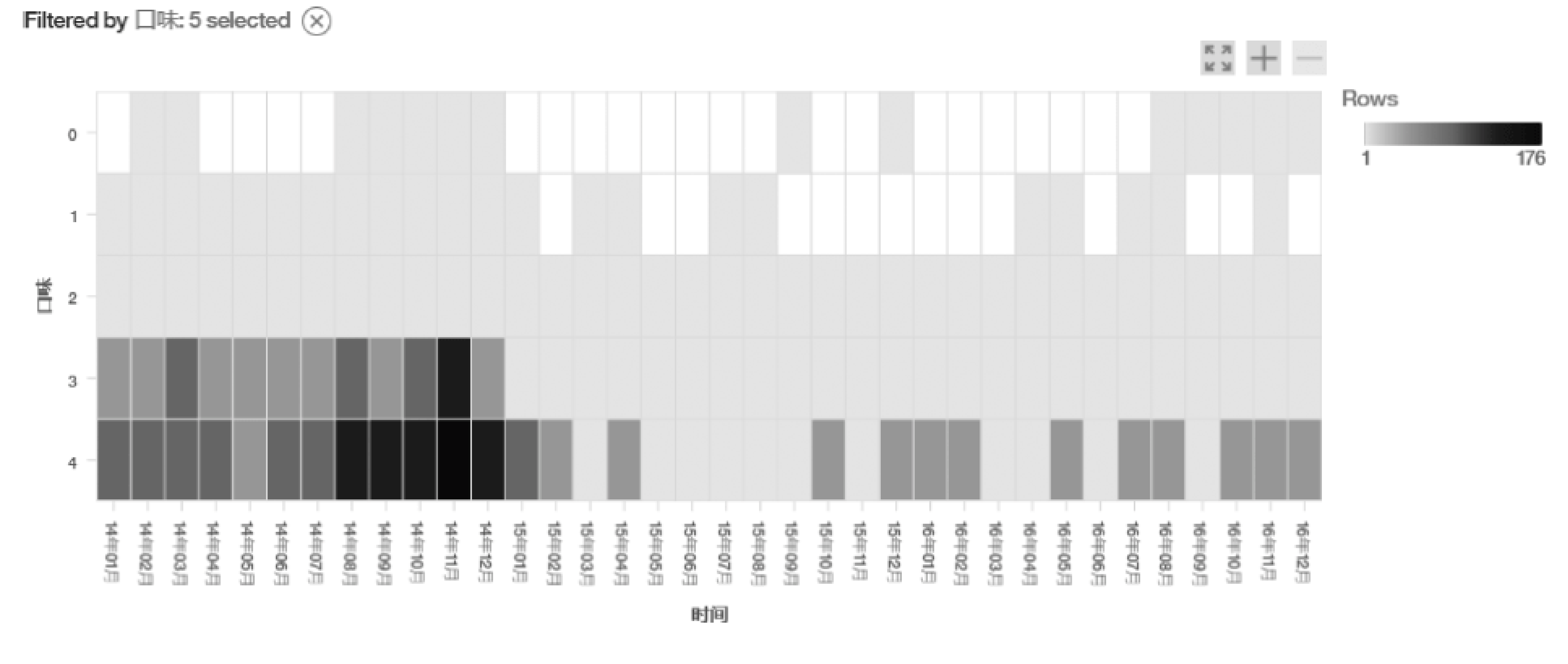


图 7.14　口味以及时间对应数量图

2014—2016 年这段时间是什么导致了该店的销量下降。在“大众点评”上有商家信息的北京海淀区北太平庄地段共有火锅店 48 家，在 2014 年之后出现的店铺共有 32 家，其中井格老灶、四川老巷子、宽板凳老灶火锅、沸炉火锅这些店的销量较突出。在这个案例中，由于只是抓取网站上的数据，所以不能获得真实销量的数据，以一段时间内的总评论数为依据，假设销量是与总评论数正相关的，从而推测出各个火锅店销量的情况。

共抓取 20 家店的数据，图 7.15 是各店建店以来的平均月评论数(这 20 家店名依次为井格老灶、全香阳坊、口福居、四川老巷子、大得涮肉、宽板凳老灶、小牛海记、小码头、欢乐牧场、池记串吧、沸炉火锅、海底捞、牡丹园涮肉、玉林串串香、老门框、芦月轩、蒸汽石锅鱼、虾吃虾涮、雪中鲜渔村、黔道贵州)。

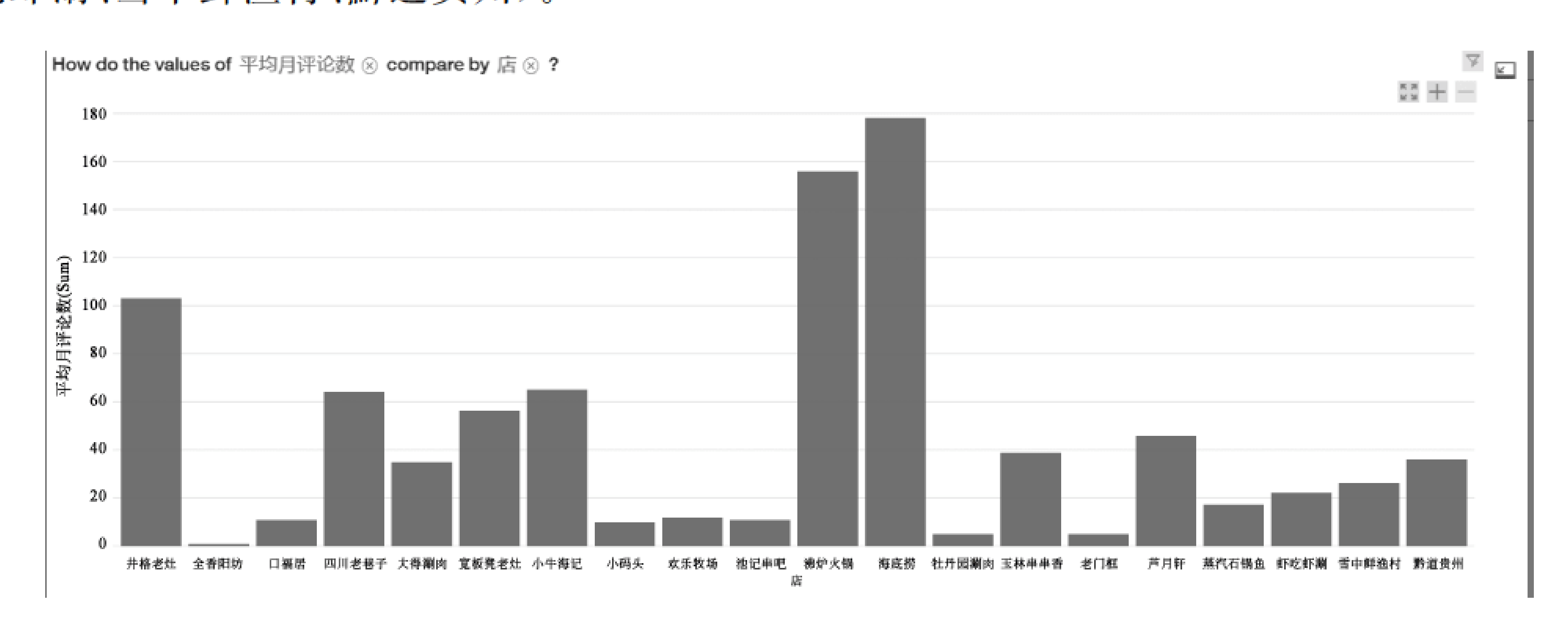


图 7.15 各店自建店以来平均月评论数

其中，海底捞、沸炉火锅两家店的平均月评论数明显多于其他店铺，而图 7.16 显示了各店 2016 年的总评论数。

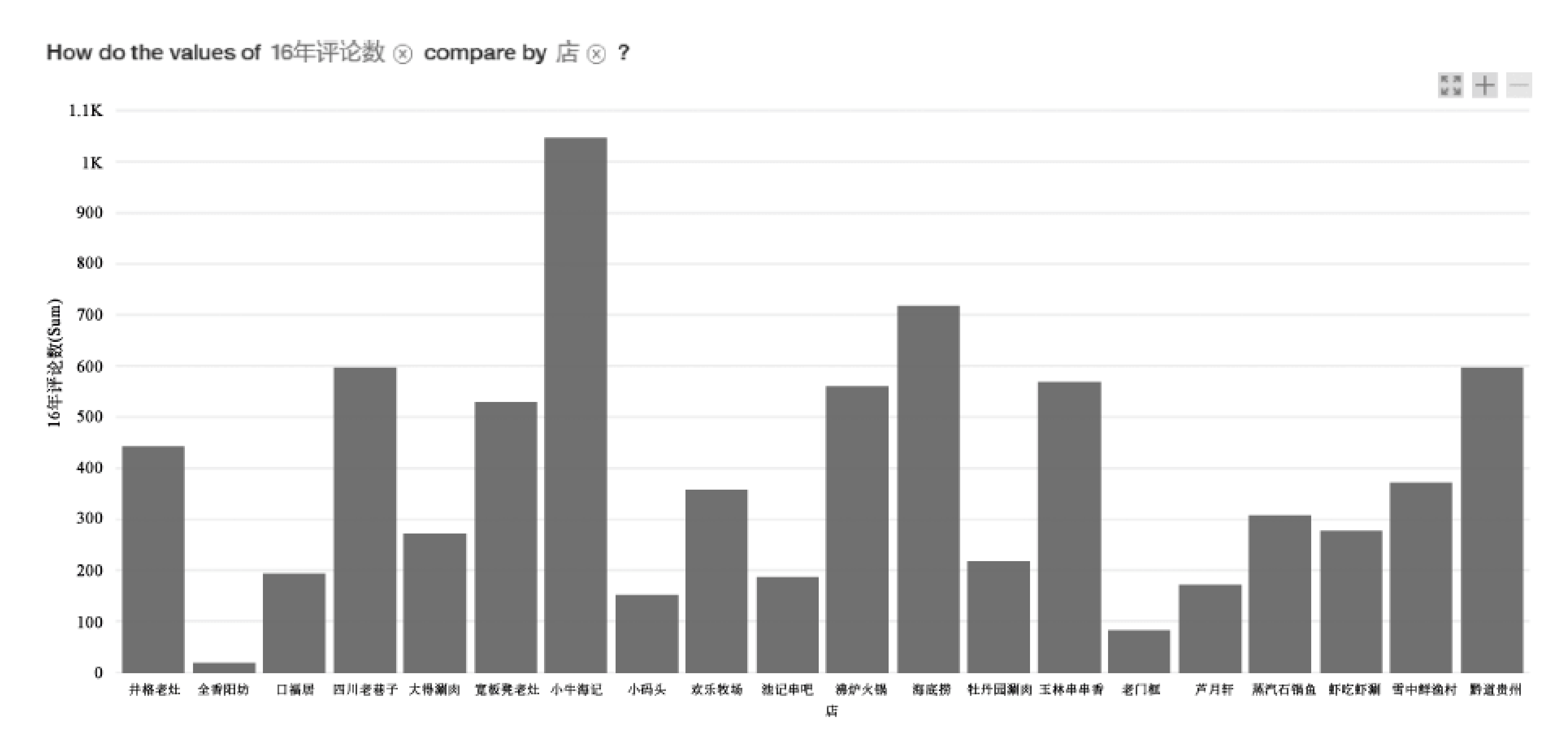


图 7.16 各店 2016 年的总评论数

可以很明显地看到小牛海记在 2016 年的总评论数遥遥领先，而海底捞店的总评论数虽然位居第二位，但与一些其他火锅店(如四川老巷子、黔道贵州等)的差距却并不大。再结合之前海底捞 2014 年的评论数与 2015 年和 2016 年的对比可以发现，海底捞(牡丹园店)的竞争力已经大不如前，且正处在一个下降期。也就是说，从总的顾客数量来讲，在这个地段并

没有显著减少。只不过，海底捞(牡丹园店)的一部分顾客被其他的该地段火锅店瓜分了。如果考虑再开设一家新的火锅店，一定要选在一个火锅店铺相对不那么密集的地段，以减少竞争，同时又要兼顾到交通、人流量等因素，确保有一定的消费人群。

7.3.2 店铺选址分析

为了避免火锅业盲目跟风现象的出现，如果海底捞想要开设新的分店，需要考虑上面提到的交通、人流量、竞争以及自身的因素，为了更好地分析和选择开店位置，可以抓取海淀区的其他行政区的店铺信息以及一些热门行政区的店铺信息。因为需要规避开其他的海底捞火锅店，所以抓取的都是尚未开设海底捞火锅店的区域中前三位的店铺信息。抓取得到的数据处理后如图 7.17 所示。

	A	B	C	D	E	F	G	H	I	J	K
1	店名	所在区域	16年总评论数	口味	环境	服务	人均	团购	外送	总评论数	区域总评论
2	黄记煌三汁焖锅(圆明园百旺店)	农业大学西区	168	7.6	7.3	7.3	77	有	无	868	2463
3	虾吃虾涮(农大店)	农业大学西区	317	9	8.4	8.4	68	有	有	818	2463
4	京都蝎府羊蝎子(农大店)	农业大学西区	398	9	8.8	8.2	79	有	有	777	2463
5	沸腾夜话重庆私房鱼锅(北蜂窝路店)	军博	237	8	7.7	7.6	82	有	无	1813	3958
6	沸炉火锅.地道的川味火锅(军事博物馆店)	军博	453	9	8.9	8.9	98	有	无	1766	3958
7	老北京兴缘斋涮羊肉(甘家口店)	军博	117	7.8	7.3	7.1	102	有	有	379	3958
8	高兴火锅(知春路店)	知春路	1488	9.1	9	9.3	99	有	无	6176	13276
9	友仁居老北京涮羊肉(知春路店)	知春路	496	8.1	7.3	7.4	76	有	无	3071	13276
10	口福居火锅(大运村店)	五道口	440	7.6	7.4	7	77	有	无	4032	13727
11	九宫格重庆火锅(五道口店)	五道口	660	8.4	8.4	8.4	101	有	无	3516	13727
12	新辣道鱼火锅(欧美汇店)	中关村	1120	8.6	8.3	7.8	95	无	有	5734	12174
13	赵香兰南洋火锅铺(欧美汇购物中心店)	中关村	1140	8	8.5	8.1	92	有	无	4405	12174
14	呷哺呷哺(新中关购物中心店)	中关村	160	7.5	6.6	6.6	47	无	无	2035	12174
15	四川仁火锅(双井店)	双井	1620	8.3	8.1	7.9	93	无	无	9007	17085
16	黄门老灶火锅(双井旗舰店)	双井	1500	9.1	8.5	8.7	125	有	无	4550	17085
17	蝎王府(垂杨柳店)	双井	1440	9	8.1	8.2	84	有	有	3528	17085
18	和合四季椰子鸡火锅(卓展店)	五棵松	2740	8.8	9	8.9	110	有	无	4070	10105
19	竹园村火锅	五棵松	560	7.8	7.1	7	84	无	无	3286	10105
20	陈阿婆重庆绿色鱼火锅(永定路店)	五棵松	580	8.5	8.4	8.1	88	有	无	2749	10105
21	蜀正园火锅	魏公村	660	8	7.8	7.7	88	有	无	2918	9174
22	重八牛府(小刀牛试店)	魏公村	3120	9.2	9.1	9.3	92	有	无	4085	9174
23	新辣道鱼火锅(华宇时尚购物中心店)	魏公村	740	8.7	8.5	7.9	87	有	无	2171	9174
24	阿丽斑鱼庄	田村	560	8.5	7.5	7.5	87	有	有	1398	2311
25	百叶居	田村	260	8.5	8.1	8.1	98	有	无	913	2311

图 7.17 各区域前三位的店铺信息

将数据导入 Watson analytics，数据评分接近 90%，为优良数据。因为 2016 年总评论数最能反映店铺在 2016 年的火爆程度，所以提出问题"What drives 2016 年总评论数"，得到图 7.18。

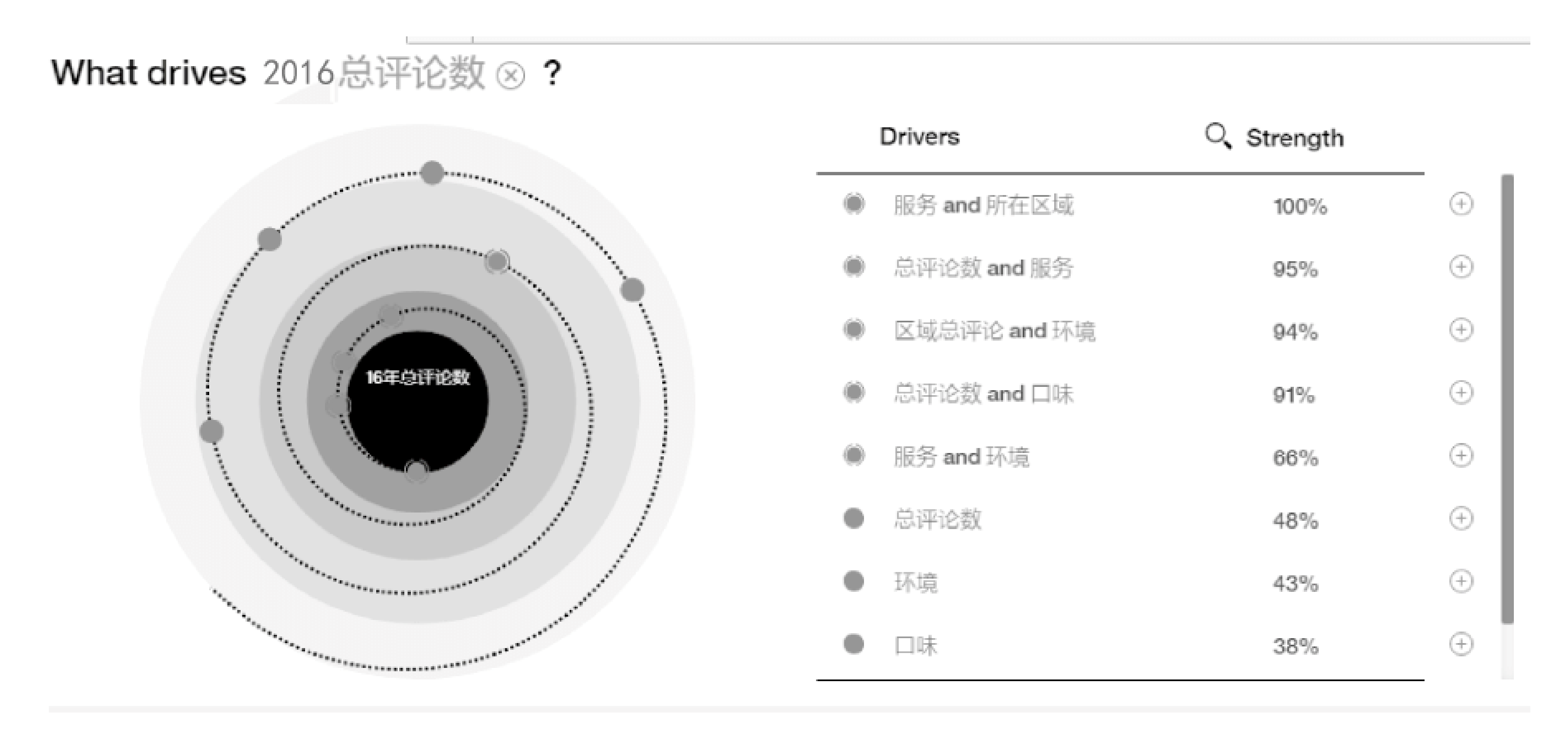


图 7.18 对于各店铺而言 2016 年总评论数的影响因素

对于火锅店来说，最关键的影响因素是服务和所在区域，其次是总评论数（对应这家店积累的口碑与顾客资源）、环境和口味，至于能否进行团购，是否配送外卖以及人均消费金额，这些并不是影响一家店受欢迎程度的主要因素。因此，分析 2016 年总评论数和所在区域的关系，提出问题“What are the values of 2016 年总评论数 for each 所在区域”，得到结果，如图 7.19 所示。



图 7.19 反映各区域人流量的总评论数图

从图 7.19 中可以看出，双井、五棵松以及魏公村是 2016 年生意最好的区域，其次是大钟寺、中关村、知春路、五道口等区域。那么，为了获取更多的客源，优先分析生意最好的区域。图 7.20 是双井区域口味、环境、服务和人均消费与 2016 年总评论数的关系。

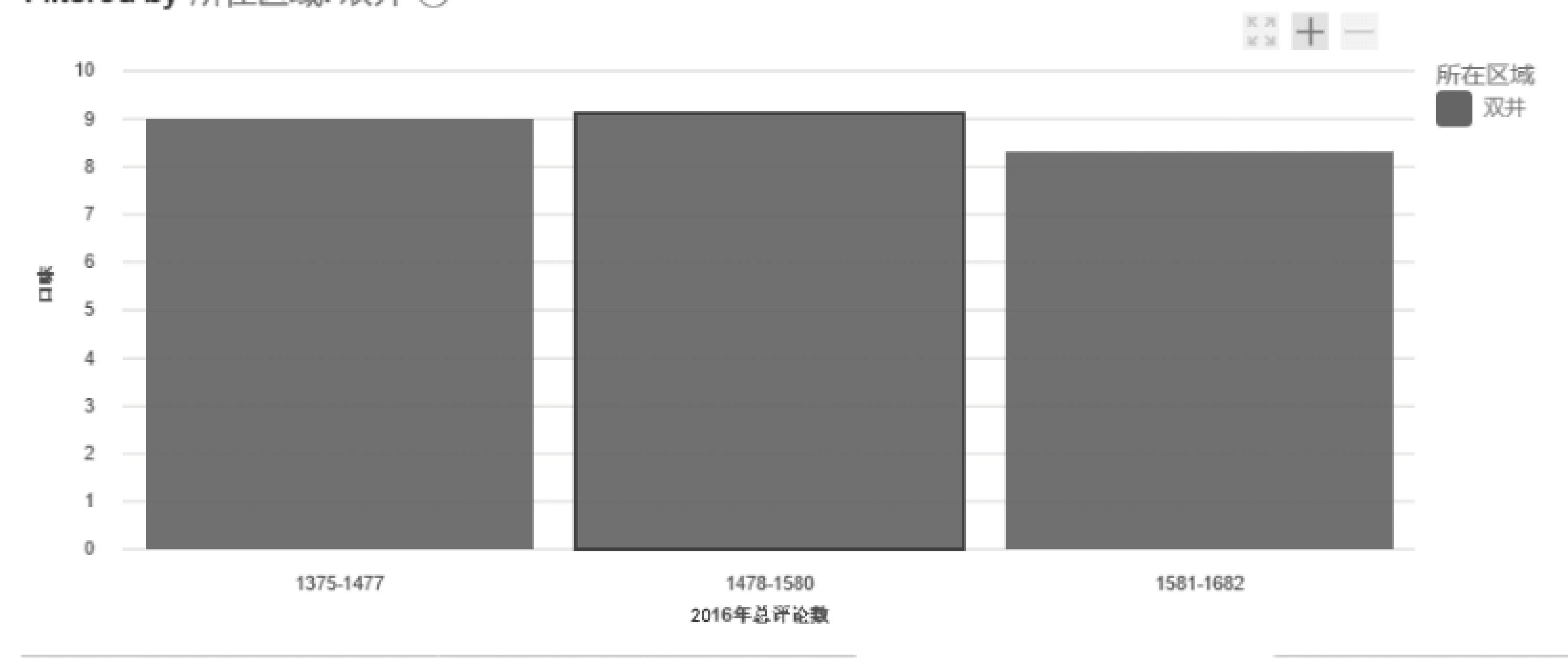
通过这些数据可以推测，双井这个区域中口味、环境、服务，尤其是人均消费适中的店反而是最受欢迎的，海底捞的服务的优势很难发挥，并且可以看到双井这里抓取的代表性的店铺的 2016 年总评论数是比较接近的，也就是说，竞争相对激烈，所以双井并不适合海底捞开设新的分店。

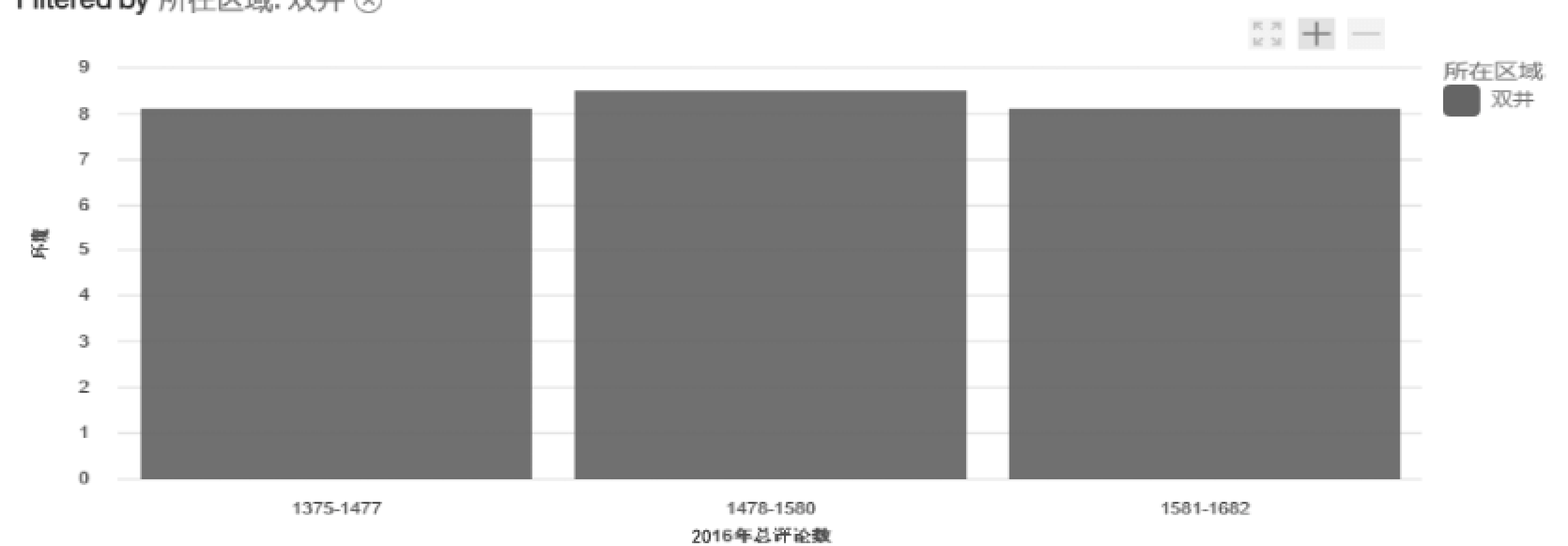
类似地，可以分析魏公村和五棵松区域。可以看到，在魏公村区域存在着一家生意火爆的火锅店——重八牛府，其他火锅店竞争不过这家火锅店，而且这家店的口味、服务以及环境都在 9 分以上，人均消费对销量的影响并不大，虽然海底捞可能会在与这家店的竞争中处于下风，但重八牛府与其他火锅店之间差距最大的地方是服务，也就是说，海底捞的优势有发挥之处，而且该地区的人流量有一定的保证，所以可以考虑在魏公村建设分店。五棵松区域中影响最大的因素是人均消费，而海底捞处于一个不占优势的人均消费区间，其次的因素是服务与环境，海底捞的环境因素也不占优势，服务因素占优势，这样的区域也不是很适合海底捞开设新的分店店铺。

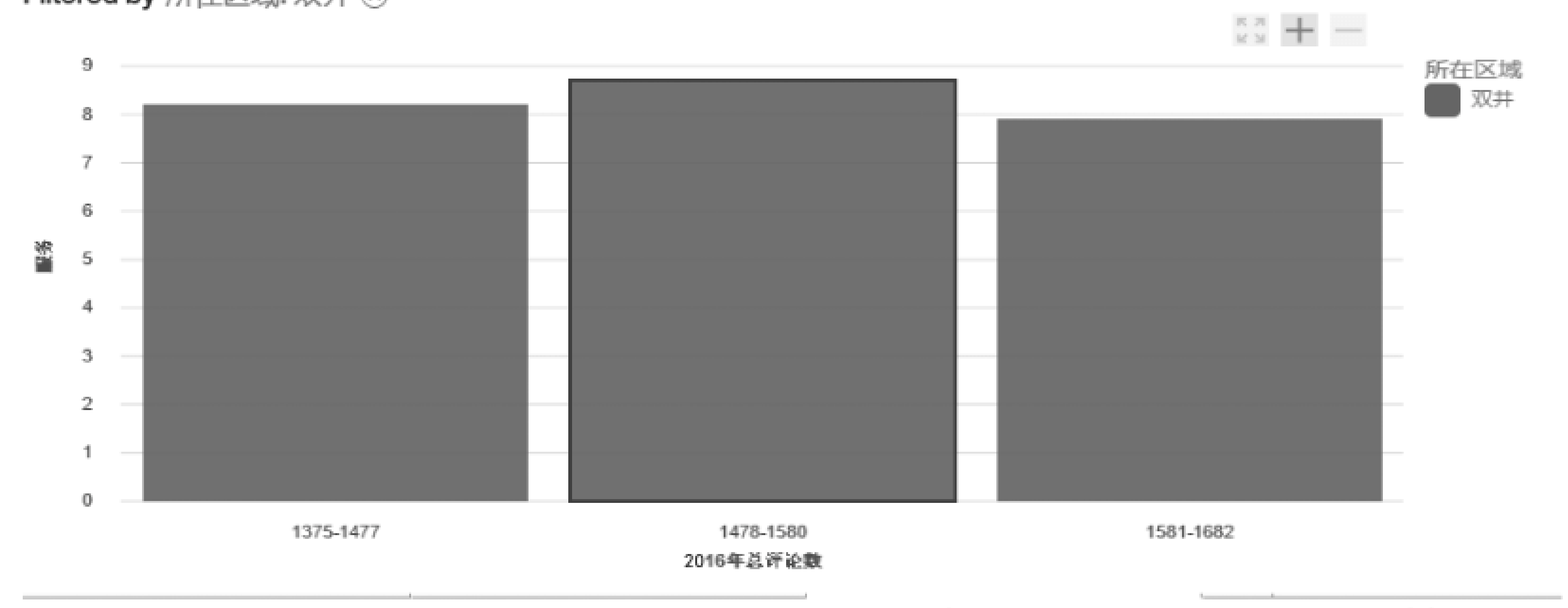
再根据区域与服务的关系分析，如图 7.21 所示。

因为这两个因素是影响最大的因素，同时海底捞的优势也在于服务，所以根据这幅图进一步分析。魏公村、西三旗、知春路、五道口、中关村中服务因素可以对销量有明显影响，比较适合海底捞这种服务方面有优势的店铺，再加上之前的对人流量较大的区域的分析，推荐在魏公村开设新的店铺。

作为一家火锅店，为了能够获取更多的利润，就需要得到更多的客户。一般来说，对于







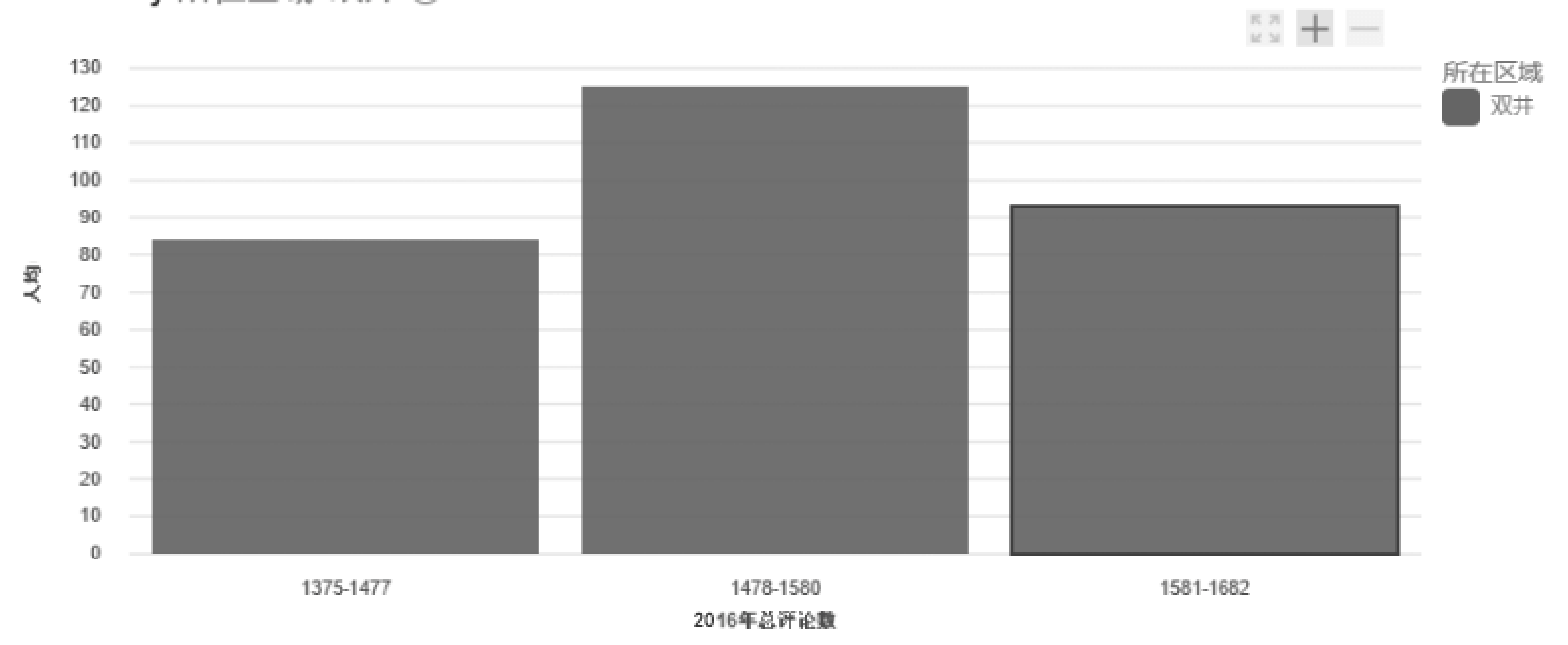


图 7.20 双井区域口味、环境、服务和人均消费与 2016 年总评论数的关系

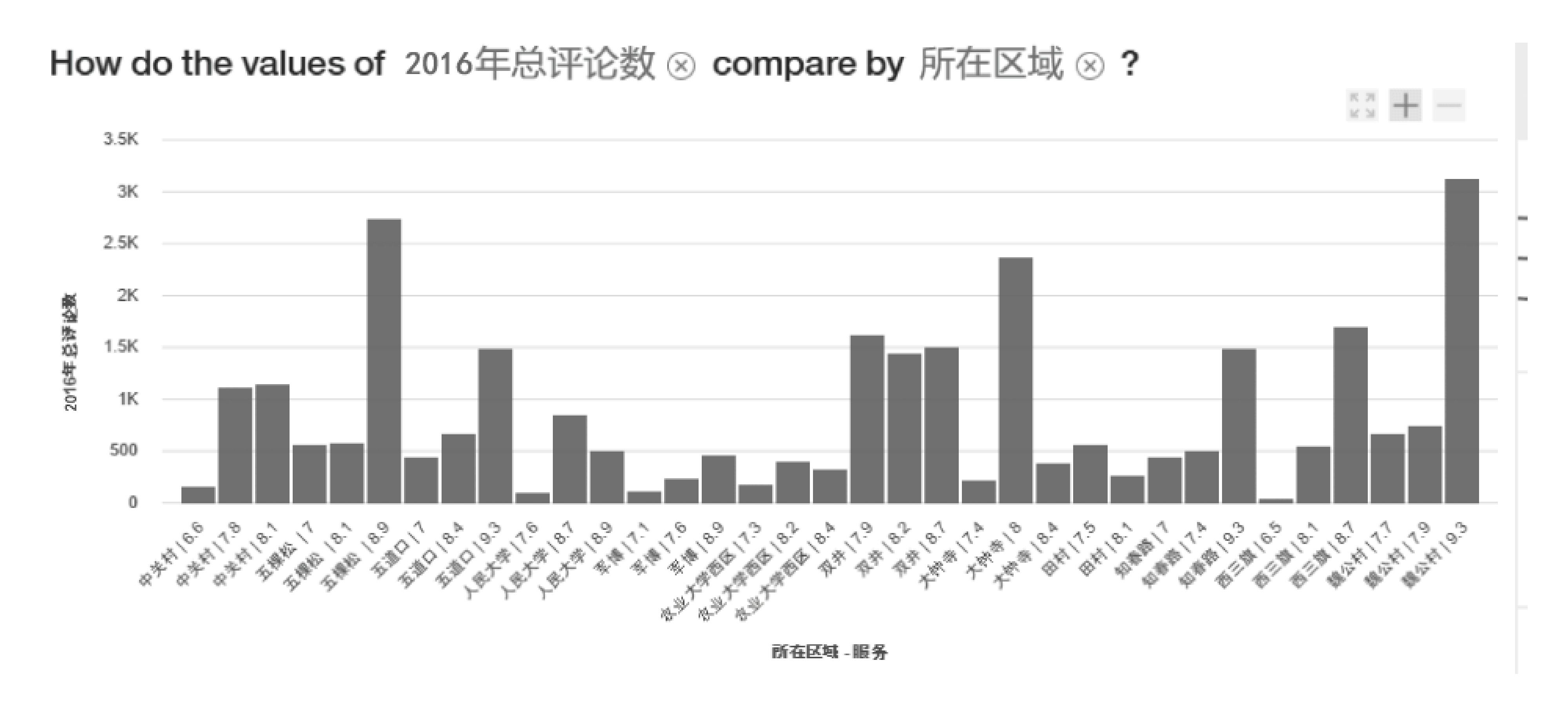


图 7.21 各区域、服务与 2016 年总评论数的关系

饮食行业来说，菜品口味是衡量店铺是否受欢迎的关键因素，所以可以结合大众点评网站上给出的推荐菜做出关于菜品的营销建议。

如图 7.22 所示，根据预处理得到的 word.txt 以及大众点评的推荐菜找到受欢迎程度较低的菜，包括简阳鱼、金针菇、香蕉酥、猪脑、牛肉丸、简州鱼、拉面以及未上榜的菜品。这些菜不那么受欢迎可能是因为这些菜不适用于火锅这种烹饪方式，也可能是因为本店的对应菜品进货源不够好，导致菜品质量存在一些问题，所以菜品没有达到应该有的受欢迎程度，还有其他的可能性，为了进一步了解，可以分析竞争对手店中的顾客喜欢的菜的情况(主要是考虑两家店中相同的菜的受欢迎程度)。此处抓取主要竞争对手之一——小牛海记潮汕牛肉店(牡丹园店)的喜欢的菜的数据，处理之后如图 7.23 所示。

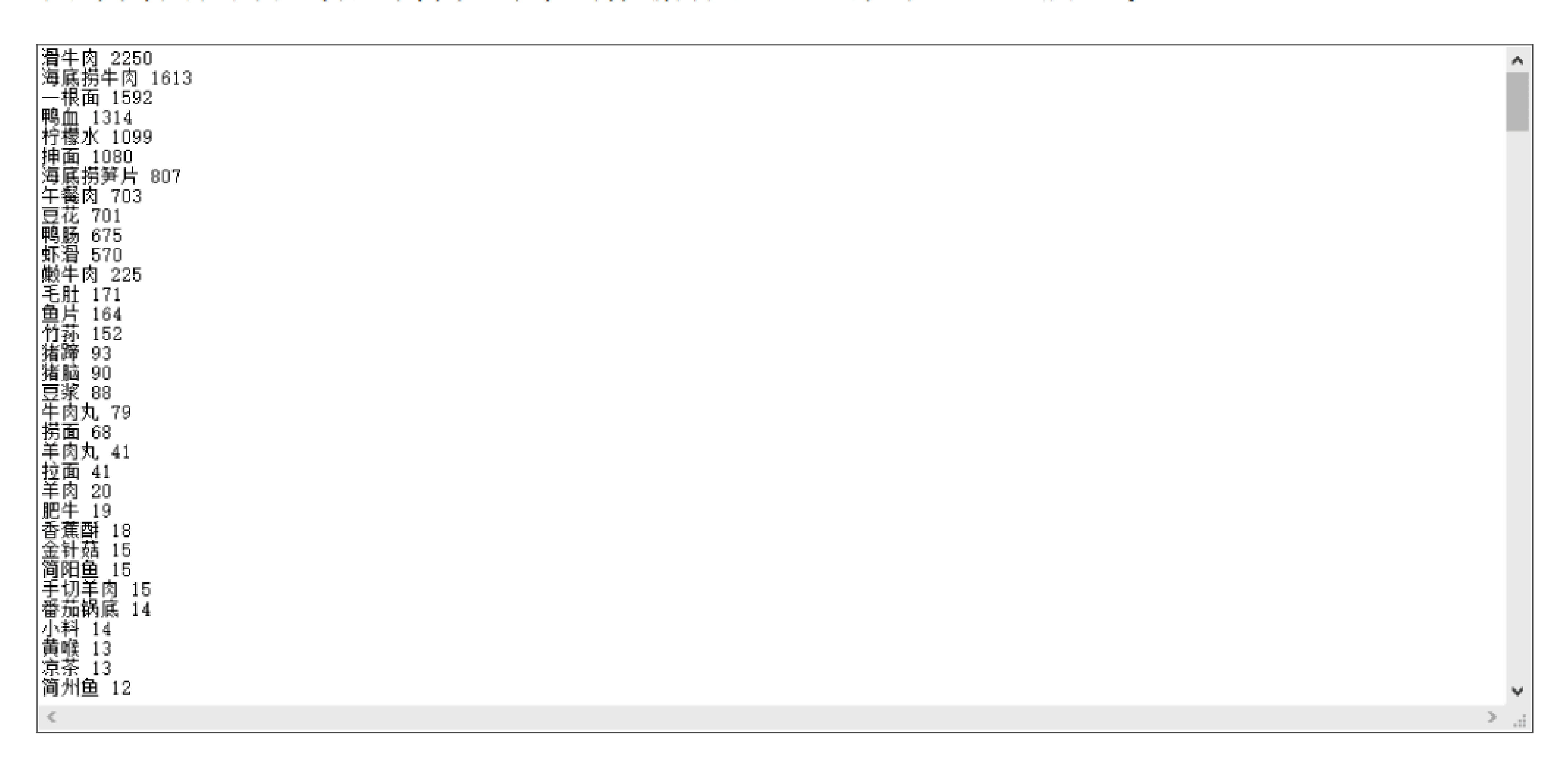

滑牛肉 2250
海底捞牛肉 1613
一根面 1592
鸭血 1314
柠檬水 1099
抻面 1080
海底捞笋片 807
午餐肉 703
豆花 701
鸭肠 675
虾滑 570
嫩牛肉 225
毛肚 171
鱼片 164
竹荪 152
猪蹄 93
猪脑 90
豆浆 88
牛肉丸 79
捞面 68
羊肉丸 41
拉面 41
羊肉 20
肥牛 19
香蕉酥 18
金针菇 15
简阳鱼 15
手切羊肉 15
番茄锅底 14
小料 14
黄喉 13
凉茶 13
简州鱼 12

图 7.22 菜品及受欢迎程度

同样要结合大众点评网站上给出的推荐菜情况。对比分析，发现金针菇以及年糕、米糕、香蕉酥这些点心类的食物的受欢迎程度在这两家店都相对较低，所以可以考虑适当减少这些种类菜品的储备量，还可以看到牛肉丸在小牛海记潮汕牛肉店的受欢迎程度要好于在海底捞火锅店的受欢迎程度，所以有可能是小牛海记潮汕牛肉店的货源更好一些，也有可能

梭边鱼 372
虾滑 226
嫩牛肉 98
三味鱼锅 79
酸梅汤 67
阿婆豆花 53
鲜鸭血 53
鸭血 53
羊肉 45
芒果汁 19
油豆皮 17
肥牛 12
龙俐鱼 11
豌豆尖 11
牛骨髓 11
银耳汤 10
鸭肠 10
红薯 7
牛肉丸 7
水晶包 6
手打牛筋丸 6
火锅凉茶 6
西洋菜，吊龙 4
虾仁黄瓜 4
爆米花 4
吊龙 4
兔头 4
自助小料 4
豆花 3
牛滑 2
土豆 2
红糖饼 2
炸腐竹 2
特级肥牛 2
蟹黄蟹籽丸 2
饺子 2
猪肉香菇 2
五花趾 1
虾划 1
鲜腰片 1
海三鲜 1
小凉菜 1
红枣汁 1
猪肉豆角饺子 1
鱼 1
手擀面 1
蘑菇 1
脖仁 1
菌汤锅 1
锅底 1
猪肉尖椒饺子 1
潮汕炸豆皮 1
凉茶 1
大头鱼 1
福华肥牛 1
丝瓜 1
圆生菜 1
梭 1
土豆泥茄子 1
峨眉笋 1
生菜 1
鮰鱼 1
回鱼 1
毛血旺 1
蓝莓山药 1
特价羊肉 1
熟梭边鱼 1
麻辣回鱼 1
紫薯饼 1
无渣鸳鸯锅 1
毛肚 1
西瓜 1
蔬菜手擀面 1
羊肉丸 1
麻辣兔头 1
金针菇 1
豌豆尖汤 1
抻面 1
蟹肉蒸饺 1
白菜 1
菇类拼盘 1
三鲜蒸饺 1
年糕 1
陈阿婆米糕 1
长豆角茄子饭 1
麻酱 1
鮰鱼锅底 1
酱棒骨 1
鲜虾滑 1
香菜羊肉丸子 1
素三鲜 1
绿毛肚 1

图 7.23　小牛海记潮汕牛肉店菜品受欢迎情况

是这家店的牛肉丸调味处理的方式更美味。为了深入分析，在小牛海记潮汕牛肉店的推荐菜中单击后弹出的推荐评论页面中抓取有关牛肉丸的推荐的评论。

抓取小牛海记潮汕牛肉店的有关牛肉丸推荐的评论使用的脚本与之前使用的抓取数据的脚本类似，只需要根据网页的具体的 URL 以及页面标签对代码做出部分修改。根据抓取到的内容提取词频，绘制标签云图如图 7.24 所示。

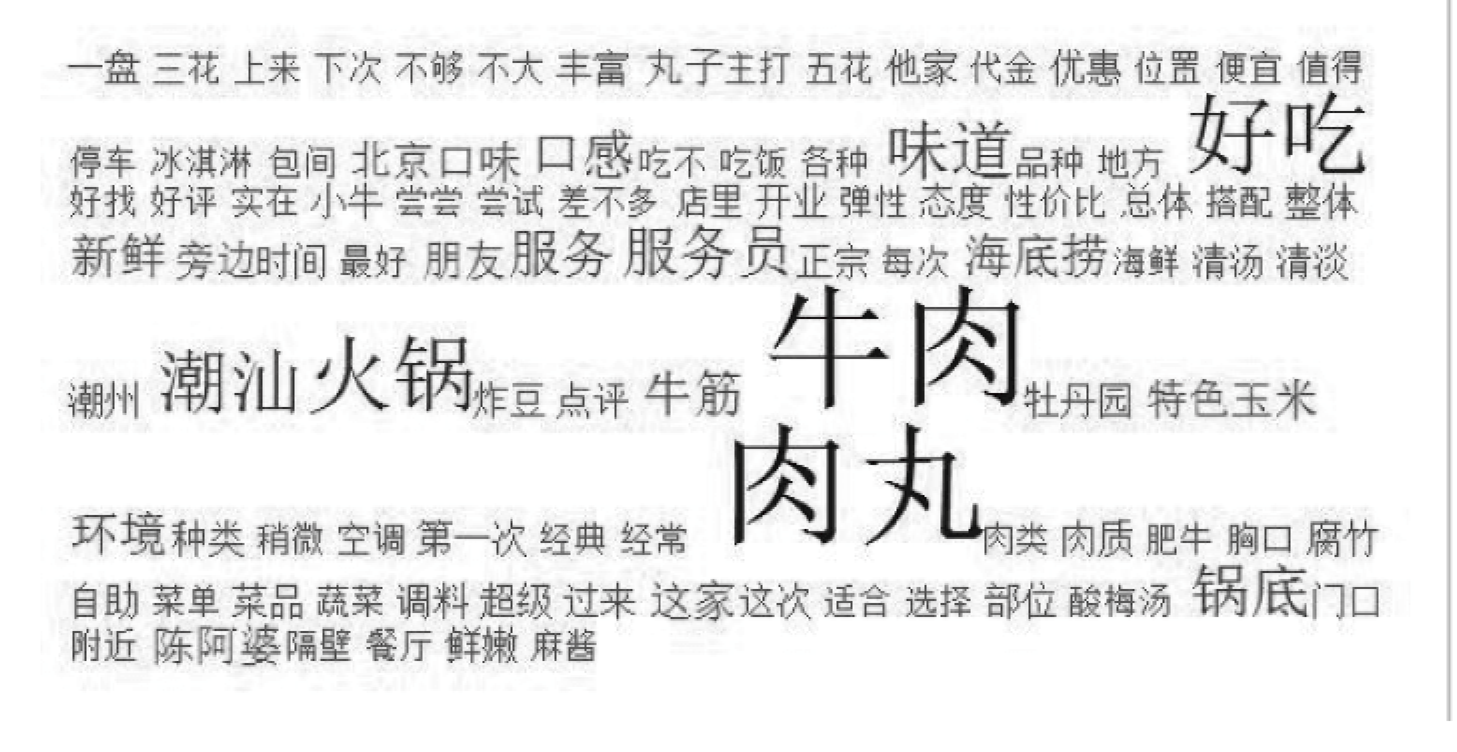


图 7.24　小牛海记潮汕牛肉店评论标签云图

首先，可以忽略牛肉、肉丸这两个不能够展示顾客感受的词汇，之后发现潮汕火锅、锅底、味道、口感这几个词的词频较高，可以据此推测，这家店的牛肉丸和店内的一些锅底十分搭配，而且牛肉丸的味道和口感都很好，海底捞如果也想让自己店内的牛肉丸更受欢迎，可以考虑增加与之配套的锅底，也可以考虑引进口感更好的牛肉丸。

此外，还可以看到在海底捞店中，简州鱼以及简阳鱼相对不那么受欢迎，反观小牛海记潮汕牛肉店中最受欢迎的就是梭边鱼，那么海底捞也可以考虑更换店内鱼类的品种，例如将简州鱼换成梭边鱼。从菜品的角度出发，还可以考虑菜品之间的相关性，分析各种菜品之间的相关性，从而更好地做出菜品推荐。

7.4　菜品关联分析

根据大众点评网站上用户填写的喜欢菜的信息进行菜品关联分析，如图 7.25 所示。

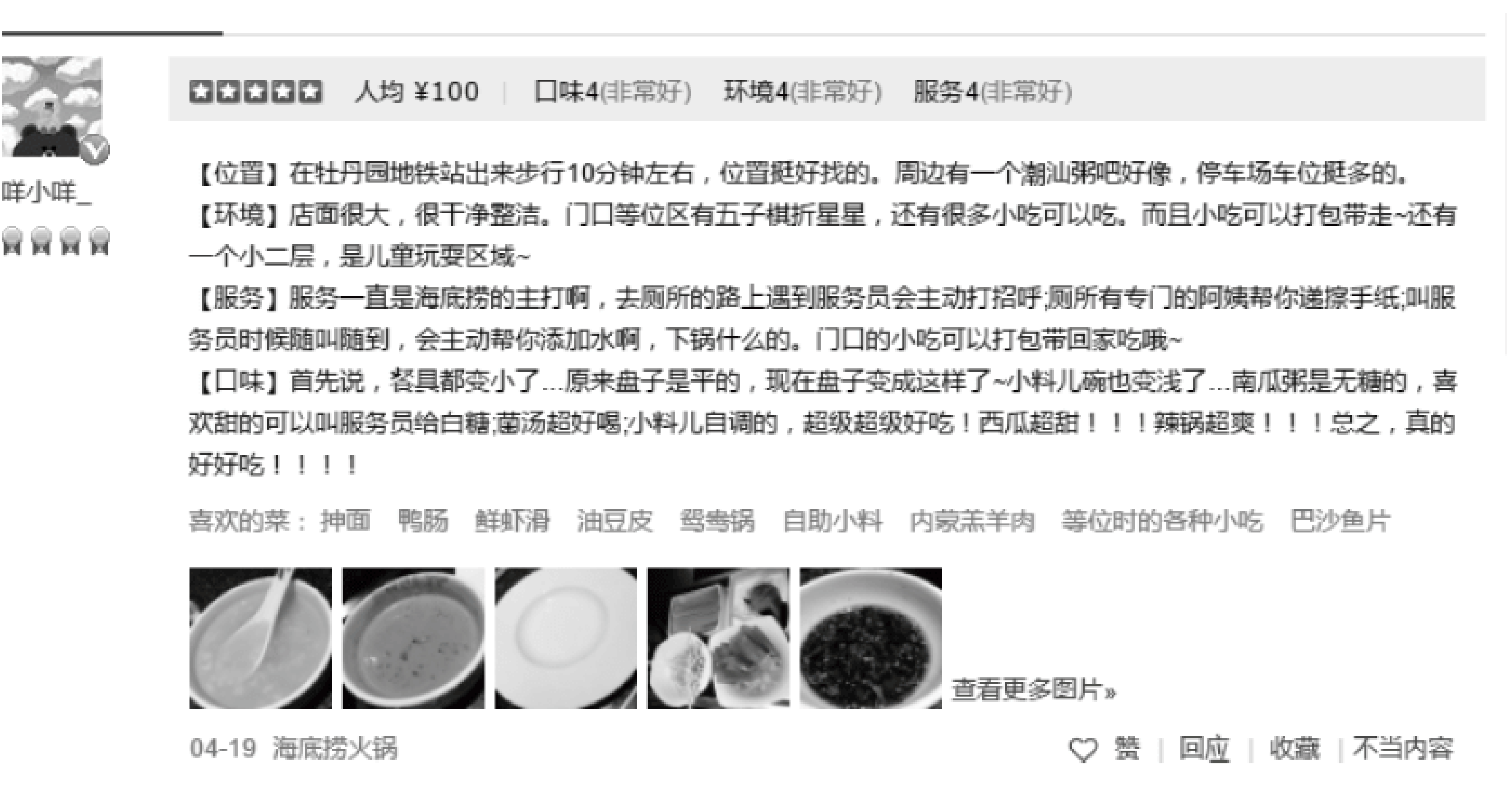


图 7.25　大众点评网站截图

菜品关联分析的目的是找到用户的推荐菜之间的关系。在抓取的数据中，每个用户的每个推荐菜都分布在一个单元格中，没有统一的列名称，这样的数据导入 SPSS Modeler 后，会因为没有统一列名称而无法进行分析。需要将所有的菜品统计成为列名称，并看每一个用户的推荐菜是否出现在列名称中，若出现，则设为 1，反之设为 0。

对数据进一步预处理利用到前面的 word.txt 和“菜品.xlsx”，用 Python 实现。选取推荐数大于 10 的菜进行关联分析，将每个菜名设置为新的表格的列名称，并读取抓取的“菜品.xlsx”中的每一用户的推荐菜，若列名称中的菜出现在该用户的推荐菜中，则将对应的单元格设为 1，否则设为 0。

```
#!/usr/bin/python3
import xlwt
import xlrd

# 要输出的表格
workbook = xlwt.Workbook()
sheet1 = workbook.add_sheet('sheet1',cell_overwrite_ok = True)

# 读取统计的词频
f = open('./word.txt','r',encoding = 'UTF-8')
content = f.readlines()
f.close()
```

```
# 添加第一行的菜品名
num = 0
writeNum = 0
mealTup = ()                                    # 当前输出的所有菜品
while num < len(content):
    tem = content[num].find(' ')
    mealCount = content[num][tem: - 1]          # 菜品数量
    if int(mealCount) >= 10:                    # 选取数量大于 10 的菜
        content[num] = content[num][0:tem]
        sheet1.write(0,writeNum,content[num])
        mealTup = mealTup + (content[num],)
        writeNum += 1
    num += 1

# 读取抓取的数据
workbook1 = xlrd.open_workbook('./菜品.xlsx')
worksheets = workbook1.sheet_names()
worksheet1 = workbook1.sheet_by_name(u'其余的评价')

num_rows = worksheet1.nrows
for curr_row in range(num_rows):
    row = worksheet1.row_values(curr_row)       # 每一行

    mealFlag = 0
    while mealFlag < len(mealTup):
        mealName = mealTup[mealFlag]
        try:
            row.index(mealName)
            sheet1.write(curr_row + 1,mealFlag,1)   # 该用户有推荐,设为 1
        except:
            sheet1.write(curr_row + 1,mealFlag,0)
        mealFlag += 1
workbook.save('meal.xls')
```

处理后的 meal.xls 如图 7.26 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	滑牛肉	海底捞牛肉	一根面	鸭血	柠檬水	抻面	海底捞笋片	午餐肉	豆花	鸭肠	虾滑	嫩牛肉	毛肚
2	1	0	0	1	1	0	0	0	0	0	0	0	0
3	1	0	1	0	0	0	1	1	0	1	1	0	0
4	0	0	0	1	0	0	0	0	0	0	1	1	0
5	0	0	0	0	0	0	0	0	0	1	0	0	0
6	1	0	1	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	1	0	1
9	1	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	0	0	1	0	1
11	0	0	0	0	0	0	0	0	0	1	0	0	1
12	0	0	1	0	0	0	0	0	0	0	1	1	0
13	1	1	1	0	0	0	1	0	0	0	1	1	0
14	1	1	1	1	0	0	0	0	0	0	0	0	1
15	0	0	1	0	0	0	0	0	0	0	1	1	0
16	0	0	0	0	0	1	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	1	1	0	0	0	0	0	0	0	0	0
19	1	0	0	1	0	0	0	0	0	0	0	1	0
20	0	0	0	0	0	0	0	0	0	0	1	0	0
21	0	1	0	0	1	0	0	0	0	1	1	0	1
22	1	1	1	1	0	0	0	0	0	1	1	0	0

图 7.26 Excel 离散化后的数据

然后将该表格导入 SPSS Modeler 18.0，单击“插入”→“源”→Excel，选择文件类型和导入文件。设置数据类型为“分类”，这里只有“0”和“1”两种值，所以也可以设为“标志”。添加“类型”节点，这里需要比较所有字段间的关系，所以将所有字段的角色都设为“任意”，如图 7.27 所示。



图 7.27　设置“类型”节点

单击“插入”→“建模”→Apriori，设置 Apriori 节点中的“最低条件支持度”为 5.0，“最小规则置信度”为 10.0，单击“运行”得到结果模型。双击模型查看，如图 7.28 所示。按照“支持度”排序，“支持度”指含有前后项的记录在总体中的占比，可以看出推荐菜的排名。“滑牛肉”“海底捞牛肉”和“一根面”最受欢迎。其次选择按照“规则支持”排序，“规则支持”指的是前项和后项同时出现的记录在总体的占比。这里商家可以针对那些经常一起出现的菜，设置一些菜的套餐，例如，可以推出“滑牛肉”“海底捞牛肉”和“一根面”3 个菜的组合菜，因为 3 个菜中任意两个都高频地同时出现。也可以推出“滑牛肉”和“鸭血”的组合菜，因为这两个同时出现的概率达到 18.628%。商家同样可以根据“规则支持”和“置信度”的排序进行菜品的相关推荐。例如，在用户点了“柠檬水”但未点“滑牛肉”的时候，可以显示“点了柠檬水的用户有 58.781%也点了滑牛肉”；在点了“海底捞笋片”而未点“滑牛肉”的时候，可以显示“有 57.993%的用户还点了滑牛肉”。通过在用户点菜的时候进行关联推荐，增加相关菜品的销售量。

根据建模的结果，将相同的前项综合在一起，进一步进行数据预处理。单击建模结果中的“将模型复制到剪贴板”，粘贴到 result.txt，将所有空白区域替换为单个空格，然后利用 Python 处理该文本文件，将相同前项的所有后项聚集在一起，结果保存到 word1.txt 中。代码如下：

```
#!/usr/bin/python3

f = open('./result.txt','r',encoding = 'UTF-8')
content = f.readlines()
f.close()
f1 = open("word1.txt",'w',encoding = 'UTF-8')
```

```
f1.write("前项 -> 后项\n")

keyDicts = {}
num = 0
while num < len(content):
    tem = content[num].split(' ')
    try:
        key = keyDicts[tem[1]]
        keyDicts[tem[1]] = key + [tem[0]]
    except:
        keyDicts[tem[1]] = [tem[0]]
    num += 1

keys = keyDicts.keys()
for key in keys:
    f1.write(str(key) + " -> " + str(keyDicts[key]) + "\n")

f1.close()
```

按以下内容进行排序：置信度 % 597 属于 597

后项	前项	支持度 %	置信度 %	规则支持 %
滑牛肉	豆花	16.698	60.342	10.076
海底捞牛肉	柠檬水 鸭血 滑牛肉	6.479	60.294	3.907
滑牛肉	午餐肉 海底捞笋片	5.336	60.268	3.216
海底捞牛肉	海底捞笋片 滑牛肉	11.148	59.615	6.646
滑牛肉	鸭血	31.301	59.513	18.628
海底捞牛肉	豆花 抻面	5.765	59.091	3.406
滑牛肉	海底捞牛肉	38.423	59.02	22.677
滑牛肉	柠檬水	26.179	58.781	15.388
海底捞牛肉	海底捞笋片 鸭血	7.646	58.567	4.478
一根面	柠檬水 鸭血 滑牛肉	6.479	58.456	3.788
海底捞牛肉	海底捞笋片 一根面	8.171	58.309	4.764
一根面	午餐肉 海底捞牛肉 滑牛肉	5.765	58.264	3.359
海底捞牛肉	海底捞笋片 抻面	6.432	58.148	3.74
海底捞牛肉	抻面 一根面	9.076	58.005	5.264

图 7.28 菜品关联挖掘结果

处理的部分结果如图 7.29 所示。“->”左边是前项，后边是后项集合。后项集合中每一项为一个后项，包括了后项的名称和置信度。商家可以根据这个整理后的数据，直接在用户选择某一个菜品的时候，出现所有与之相关的其他菜品。

网站上提供的可供用户填写的“喜欢的菜”这个模块能够为菜品的推荐提供一些数据，此外，还可以从用户正面评论中获取关于用户喜欢的菜的数据。对评论数据做一些处理，分析的菜品包括一些受欢迎程度较高的菜品。

```
前项 → 后项
滑牛肉 → ['虾滑-10.54', '鸭肠-16.81', '豆花-18.82', '午餐肉-17.48', '海底捞笋片-20.82'
海底捞牛肉 → ['鸭肠-16.80', '豆花-20.58', '午餐肉-21.82', '海底捞笋片-26.60', '抻面-29.
一根面 → ['虾滑-10.87', '鸭肠-14.20', '豆花-19.03', '午餐肉-19.97', '海底捞笋片-21.55'
鸭血 → ['虾滑-10.50', '鸭肠-22.37', '豆花-20.02', '午餐肉-21.00', '海底捞笋片-24.43',
柠檬水 → ['虾滑-11.83', '鸭肠-17.38', '豆花-21.93', '午餐肉-22.11', '海底捞笋片-24.02'
抻面 → ['虾滑-10.56', '鸭肠-17.31', '豆花-22.41', '午餐肉-22.50', '海底捞笋片-25.00',
海底捞牛肉+滑牛肉 → ['鸭肠-18.59', '豆花-24.37', '午餐肉-25.42', '海底捞笋片-29.31',
一根面+滑牛肉 → ['鸭肠-14.99', '豆花-21.66', '午餐肉-22.43', '海底捞笋片-24.62', '抻面-
海底捞笋片 → ['鸭肠-18.84', '豆花-26.77', '午餐肉-27.76', '抻面-33.46', '柠檬水-32.71'
鸭血+滑牛肉 → ['鸭肠-23.15', '豆花-22.25', '午餐肉-22.89', '海底捞笋片-27.37', '抻面-32
午餐肉 → ['虾滑-11.10', '鸭肠-22.05', '豆花-23.04', '海底捞笋片-31.86', '抻面-34.57',
豆花 → ['鸭肠-17.83', '午餐肉-23.11', '海底捞笋片-30.81', '抻面-34.52', '柠檬水-34.38'
鸭肠 → ['虾滑-17.04', '豆花-18.52', '午餐肉-22.96', '海底捞笋片-22.52', '抻面-27.70',
一根面+海底捞牛肉 → ['鸭肠-15.81', '豆花-25.08', '午餐肉-27.96', '海底捞笋片-30.40',
柠檬水+滑牛肉 → ['鸭肠-17.49', '豆花-25.08', '午餐肉-24.30', '海底捞笋片-26.01', '抻面-
抻面+滑牛肉 → ['鸭肠-18.23', '豆花-26.76', '午餐肉-24.58', '海底捞笋片-28.43', '柠檬水-
虾滑 → ['竹荪-10.18', '毛肚-15.26', '嫩牛肉-12.63', '鸭肠-20.18', '豆花-11.23', '午餐
鸭血+海底捞牛肉 → ['鸭肠-23.98', '豆花-26.58', '午餐肉-29.00', '海底捞笋片-34.94', '抻面
鸭血+一根面 → ['鸭肠-20.42', '豆花-23.82', '午餐肉-25.90', '海底捞笋片-28.54', '抻面-30
柠檬水+一根面 → ['鸭肠-17.05', '豆花-25.68', '午餐肉-27.79', '海底捞笋片-29.26', '抻面-
```

图 7.29　后项聚集数据

```
all_food = ['虾滑','一根面','滑牛肉','海底捞牛肉','鸭肠','嫩牛肉','毛肚','柠檬水',
'抻面','海底捞笋片','鱼片','午餐肉','豆花','豆浆','鸭血','牛肉丸','捞面','猪脑',
'猪蹄','番茄锅底','羊肉丸','鲜虾滑','肥牛','金针菇','小料','鱼豆腐','豆皮',
'简阳鱼','黄喉','肥肠','手切羊肉','竹荪','海底捞小料','冻豆腐','鸭舌','墨鱼滑',
'豌豆尖','免费水果','千层肚','小吃','鸳鸯锅','牛蛙','蒿子秆']
```

所使用的评论为抓取到的全部评论。首先从这些评论中为每个菜品找到有关的评论。判断评论是否符合标准的规则为：如果某条评论中提到了某种菜品，则将该条评论视为菜品的相关评论，添加到菜品的评论列表中，之后再写入文件。具体代码如下：

```
for x in range(1,len(all_food) + 1):
    comment = codecs.open('数据/comments.txt','r','utf-8')
    filee = open("菜品/" + all_food[x-1] + ".txt",'w')
    while 1:
        line = comment.readline()
        if not line:
            break
        if line.find(all_food[x-1])!= -1:
            filee.write(line)
```

然后将得到的各自菜品对应的评论分别存储在一个 txt 文件中，利用结巴分词对得到的评论做分词处理。结巴分词是一个 Python 中文分词组件，支持 3 种分词模式：精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适用于搜索引擎分词。同时还支持自定义词库。使用结巴分词需要先安装对应的库，安装指令为 pip install jieba 或 pip3 install jieba，代码对于 Python2/3 是兼容的。之后利用其中的分词的方法做分词以及词频统计，并且存储到菜品各自对应的文件中。具体代码如下(jiebaD.py)：

```
import jieba
```

```
def fenci(filename) :
    f = open("菜品/" + filename,'r + ')
    file_list = f.read()
    f.close()

    seg_list = jieba.cut(file_list,cut_all = True)

    tf = {}
    for seg in seg_list :
        #print seg
        seg = ''.join(seg.split())
        if (seg != '' and seg != "\n" and seg != "\n\n") :
            if seg in tf :
                tf[seg] += 1
            else :
                tf[seg] = 1

    f = open("菜品处理/result_" + filename,"w + ")
    for item in tf:
        #print item
        f.write(item + " " + str(tf[item]) + "\n")
    f.close()

if __name__ == '__main__':
    all_food = ['虾滑','一根面','滑牛肉','海底捞牛肉','鸭肠','嫩牛肉','毛肚','柠檬水',
'抻面','海底捞笋片','鱼片','午餐肉','豆花','豆浆','鸭血','牛肉丸','捞面','猪脑',
'猪蹄','番茄锅底','羊肉丸','鲜虾滑','肥牛','金针菇','小料','鱼豆腐','豆皮',
'简阳鱼','黄喉','肥肠','手切羊肉','竹荪','海底捞小料','冻豆腐','鸭舌','墨鱼滑',
'豌豆尖','免费水果','千层肚','小吃','鸳鸯锅','牛蛙','蒿子秆']
    for x in range(1,len(all_food) + 1):
        jieba.add_word(all_food[x - 1])
    for x in range(1,len(all_food) + 1):
        fenci(all_food[x - 1] + ".txt")
```

jieba.cut 方法为分词的方法，此处选择的是全模式，jieba.add_word 方法为自定义词库向词典中添加词条的方法，这里把菜品的词汇一一添入，此外还可以统计各词汇所占的权重，代码如下(jiebaE.py)。其中，jieba.analyse.extract_tags()方法提取出比重在前 50 的词汇，并且输出对应的占比情况到各个菜品的对应文件。

```
import jieba
import jieba.analyse
def fenci(filename) :
    f = open("菜品/" + filename,'rb')
    file_list = f.read()
    f.close()
    seg_list = jieba.analyse.extract_tags(file_list,topK = 50,withWeight = True)
    f = open("菜品 + /result_" + filename,"w")
    for seg in seg_list:
        #print item
        f.write(str(seg[0]) + " " + str(seg[1]) + "\n")
```

```
    f.close()
if __name__ == '__main__':
    all_food = ['虾滑','一根面','滑牛肉','海底捞牛肉','鸭肠','嫩牛肉','毛肚','柠檬水',
'抻面','海底捞笋片','鱼片','午餐肉','豆花','豆浆','鸭血','牛肉丸','捞面','猪脑',
'猪蹄','番茄锅底','羊肉丸','鲜虾滑','肥牛','金针菇','小料','鱼豆腐','豆皮',
'简阳鱼','黄喉','肥肠','手切羊肉','竹荪','海底捞小料','冻豆腐','鸭舌','墨鱼滑',
'豌豆尖','免费水果','千层肚','小吃','鸳鸯锅','牛蛙','蒿子秆']
    for x in range(1,len(all_food) + 1):
        jieba.add_word(all_food[x - 1])
    for x in range(1,len(all_food) + 1):
        fenci(all_food[x - 1] + ".txt")
```

通过词频统计以及词汇、词频占比分析之前根据喜欢的菜所得到的菜品关联情况，在点了海底捞牛肉的情况下，滑牛肉出现 11 次，占比为 24.5%，位列第一位，虾滑出现 7 次，占比为 1.56%，位列第三位；在点了滑牛肉的情况下，海底捞牛肉出现 11 次，占比为 2%，位列第十位，虾滑出现 66 次，占比为 12.2%，位列第一位，鸭血出现 33 次，占比为 6%，位列第二位，柠檬水出现 22 次，占比为 4%，位列第四位；在点了虾滑的情况下，滑牛肉出现 66 次，占比为 5%，位列第一位(上述所述位列第几位为词频占比在所有菜品词频占比中的排名，之所以菜品词频占比不高，是因为有海底捞，好吃，不错等与菜品无关词汇占比较高)。

通过上述统计可以发现，通过"喜欢的菜"以及评论做出的菜品关联大致是吻合的，所以基于评论中反映的情况，依据在点了菜品 1 的情况下，对于菜品 2 的购买量这一指标设计并实现一个推荐算法，可以根据顾客输入的菜品推荐 1～3 个菜品。首先在命令行输入：python recommend.py，回车后程序运行，该推荐程序的具体操作流程如"Demo 程序"，如图 7.30 所示。

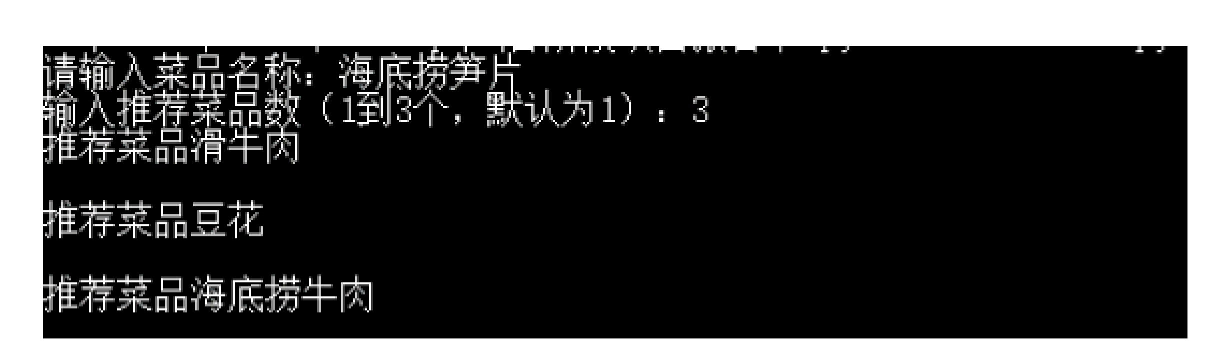


图 7.30　Demo 程序

具体代码实现(recommend.py)如下：

```
import sys
import traceback
all_food = ['虾滑','一根面','滑牛肉','海底捞牛肉','鸭肠','嫩牛肉','毛肚','柠檬水',
'抻面','海底捞笋片','鱼片','午餐肉','豆花','豆浆','鸭血','牛肉丸','捞面','猪脑',
'猪蹄','番茄锅底','羊肉丸','鲜虾滑','肥牛','金针菇','小料','鱼豆腐','豆皮',
'简阳鱼','黄喉','肥肠','手切羊肉','竹荪','海底捞小料','冻豆腐','鸭舌','墨鱼滑',
'豌豆尖','免费水果','千层肚','小吃','鸳鸯锅','牛蛙','蒿子秆']
food_name = input("请输入菜品名称：")
filename = "result_" + food_name + ".txt"
counts = input("输入推荐菜品数(1～3 个,默认为 1)：")
count = 1
try:
    count = int(counts)
except:
```

```
        print("输入错误")
        traceback.print_exc()
        sys.exit()
    count = count + 1
    temp = count
    try:
        f = open("菜品 + /" + filename,'r')
        while 1:
            line = f.readline()
            if not line:
                break
            for x in range(1,len(all_food) + 1):
                if line.find(all_food[x - 1])!= - 1:
                    count = count - 1
                    if temp!= count + 1:
                        print("推荐菜品" + all_food[x - 1] + "\n")
                    break
            if count == 0:
                break
    except:
        print("没有相关菜品推荐")
        traceback.print_exc()
```

其中的菜品和文件夹下存储的菜品是有关分词后词汇、词频占比的文件,作为推荐程序的数据支持。

除了菜品、口味,还有很多其他的影响因素影响着店铺的生意情况,为了能够进一步分析,充分利用评论中的文本内容,分析评分与评论之间存在的关联,根据评论中的内容做出更好的营销建议。

7.5 用户评论与评分的关联分析

开源的分词库"结巴分词"(https://github.com/fxsjy/jieba),对评论内容 comments.txt(预处理后得到的文件)进行关键词提取。进行关键词提取用到了基于 TF-IDF 算法的关键词抽取,TF-IDF 算法可以评估某个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。算法函数为 jieba.analyse.extract_tags(sentence,topK=20,withWeight=False,allowPOS=())。函数的接口参数中,topK 为返回几个 TF/IDF 权重最大的关键词,设置为 1000;withWeight 为是否一并返回关键词权重值,这里设置为 True。allowPOS 仅包括指定词性的词,这里设置为('n','t','s','f','v','a','b','z','m','q','x')。这里过滤掉了介词、连词、助词、叹词、代词、副词、语气词、前后缀与标点符号,保留了名词、形容词等 11 种词性。

大致浏览本店评论后,结合上文的高推荐菜品名,添加一部分自定义语料库,如"抻面""一根面"这样的食材,"棋牌""游乐场"这样的设施。最终具体的代码如下,结果保存到 commentsWord.txt 中。

```
# encoding = utf - 8
```

```
import jieba
import jieba.analyse
customizedWords = ['海底捞','滑牛肉','海底捞牛肉','一根面','鸭血','柠檬水','抻面',
                   '海底捞笋片','午餐肉','豆花','鸭肠','支付宝','微信','牡丹园',
                   '地铁站','停车位','辣锅','鸳鸯锅','儿童','面筋','会员','免费',
                   '番茄锅','变脸','外送','小吃','虾滑','嫩牛肉','毛肚','鱼片','竹荪',
                   '猪脑','捞面','香蕉酥','手切羊肉','简阳鱼','小料','黄辣丁','油豆皮',
                   '宽粉','鱼豆腐','美甲','做指甲','免费水果','排号','棋牌','表演',
                   '锅底','半份','毛巾','哈密瓜','豆浆','ipad','零食','游乐场','果盘',
                   '车位','停车']
for word in customizedWords:
    jieba.add_word(word)
with open("comments.txt",'rb') as wf,open("comemntsWord.txt",'w') as wf2:
    content = wf.read()
    words = jieba.analyse.extract_tags(content,topK=1000,withWeight=True,allowPOS=(
        'n','t','s','f','v','a','b','z','m','q','x'))
    for word in words:
        wf2.write(str(word[0]) + " " + str(word[1]) + "\n")
```

提取后的结果如图 7.31 所示，每一行为关键词及其权重。

```
海底捞 0.4420971548196067
不错 0.17279588840611504
味道 0.1551480185360137
好吃 0.12740940257270336
等位 0.12549221324548346
服务员 0.12403817283746389
火锅 0.10431640019811472
喜欢 0.08329362090090413
虾滑 0.07901559916237454
菜品 0.07321640548894003
排队 0.06821111220070897
口味 0.06591038385786242
```

图 7.31 提取后的结果

由于这样出现的结果，部分词语如“不错”“好吃”“味道”等，在实际分析的时候由于谓语或其他成分的缺失没有实用价值，所以须手动将这部分词语删除，大致删除部分词语后保留了 600 个关键词。同时需要对提取的关键词进行分组。

这里首先尝试用知识图谱的方式实现词语的自动分组。目前可用的中文知识图谱有 DBpedia、BabelNet、ConceptNet、楚辞、OpenKG.CN 和 CN-DBpedia 等几种。这里涉及大量火锅食材名，以很常见的食材名“抻面”“虾滑”和常用服务名“外卖”“半份”作为测试词汇，在上述几个知识图谱工具中进行测试，发现大部分工具的语料库都不能提取到分类名。最后选择了工具 CN-DBpedia(http://kw.fudan.edu.cn/cndbpedia/intro/)，该工具支持 RESTful 式的 API 调用。使用该工具对前文提取的关键词进行分类。Python 实现的代码如下。

```
#!/usr/bin/python3
import urllib
from urllib.request import urlopen
import json

f = open('./comemntsWord.txt','r',encoding='UTF-8')
content = f.readlines()
f.close()
f1 = open("class.txt",'w',encoding='UTF-8')

for i in range(len(content)):
    tem = content[i].split(' ')
    data = tem[0]
```

```
url_values = urllib.parse.urlencode({'entity': data})
url = "http://knowledgeworks.cn:20313/cndbpedia/api/entityAVP?"
full_url = url + url_values

reData = urlopen(full_url).read()
reData = reData.decode('UTF - 8')
reData = json.loads(reData)['av pair']

try:
    for k in range(len(reData)):
        if reData[k][0] == '分类':
            f1.write(data + '' + reData[k][1] + '\n')
except:
    pass
```

最终提取的部分结果如图 7.32 所示。

知识图谱提取的结果具有一定的参考价值，但是限于当前中文语料库和语义网的不完整，很多分类并不能自动提取出来。并且由于在分析一些词语的时候带有了主观意识，所以该部分的词语也不能正确提取，例如，“游乐区”和“婴儿车”在这里的分类是“儿童”，就完全不同于语义网中任何三元组的定义，所以知识图谱也不能提取出来。结合 class.txt 中的知识图谱提取结果，最后划分了 29 个有分析价值的组。其中包括了火锅底料、食材、小吃等食物类型；排队、服务等基础设施相关类；额外表演、外卖等附加服务类；以及顾客的类型。该分类记录在 classification.txt 中。知识图谱分类结果见表 7.1。

```
鸭血  家常菜
牛肉  肉类
抻面  汤面
鸭肠  干货
柠檬水  饮料
柠檬水  果汁
鸳鸯锅  火锅
拉面  小吃，面食
麻辣  川菜
羊肉  少数民族菜系
羊肉  家常菜
肥牛  涮食
肥牛  生食
鱼豆腐  鱼糜制品
鱼豆腐  海洋仿生食品
四川火锅  川菜
清汤  菜品的辅料
鱼片  川菜
长寿面  面食
```

图 7.32　最终提取的部分结果

表 7.1　知识图谱分类结果

分组	详细分组	详 细 内 容
料类	底料	油锅、辣椒、辣味、辣汤、宫格、麻辣锅、清汤、汤锅、麻辣、牛油、微辣、锅底、辣锅、番茄锅、鸳鸯锅、底料
	配料	配料、香油、红油、调料、酱料、小料、麻酱、芝麻、葱花、香菜、香草、清水、芝麻酱
面食	面食	抻面、一根面、拉面、面条、长寿面、捞面、杂面
荤菜	牛肉类	滑牛肉、牛滑、肥牛、牛肉、嫩牛肉、海底捞牛肉
	羊肉类	羊肉、手切羊肉、羔羊、羊肉、羊排
	其他肉类	鹅肠、鸡蛋羹、滑类、牛蛙、肥肠、猪蹄、鸭舌、脑花、黄喉、猪脑、午餐肉、肉质肉类、涮肉、肉品、丸子、鸡蛋、毛肚、鸭肠
	河鲜与海鲜	虾丸、虾滑、虾片、鱼片、沙丁鱼、墨鱼、鱼滑、泥鳅
素菜	豆制品	油豆皮、苕粉、豆腐、豆花、豆浆、鱼豆腐、皮筋、冻豆腐、豆皮
	菌类	菌类、香菇、腐竹、蘑菇、菌菇、香菇、金针菇
	笋类	竹笋、笋片、青笋、海底捞笋片、笋
	其他素菜	粉丝、茼蒿、宽粉、山药、红薯、藕片、豆苗、萝卜、蔬菜、鸭血、番茄、青菜

续表

分组	详细分组	详 细 内 容
小吃类	饮料	柠檬水、豆浆、饮料、凉茶
	水果	果盘、免费水果、柚子
	小吃	点心、油条、泡菜、蛋糕、凉菜、糍粑、小菜
	零食	冰棍、花生、小食、爆米花、烧饼、零食
出行目的	朋友聚餐	朋友、学校、室友、学生、下班、同事、同学聚会、同学
	生日	过生日、生日
	情侣	男朋友、女朋友
	家庭聚餐	一家人、全家、家庭聚会、孕妇、老人、家人
	儿童	游乐区、婴儿床、儿童、小孩子、小朋友、小孩、玩具、儿童乐园、游戏、娃娃
排队	排队	等位、拥挤、订位、排位、排队、排号、排到、排长队、等待、等待时间、等候、高峰期
服务类	服务	服务员、工作人员、服务态度、服务到位、服务生、服务质量、男服务员、服务水平、优质服务、服务员
	半份	半份
	回头客	下次、下回、多次、两次、第二次、次次、再来
额外服务	额外表演或服务	现场表演、棋牌、纸鹤、麻将、跳舞、象棋、下棋、表演、跳棋、做指甲、充电、擦鞋、贴膜、指甲、头绳、手机套、打印
	夜晚营业	晚上、半夜、夜里、夜宵
	停车	停车、停车位、停车场、车位
	外卖	外卖、外送
	团购	团购

下面将该分类结果写回 Excel。代码实现中，读取上面得到的 classification.txt，然后将每一个关键词写到新的表格文件的第一行，同时需要新建一个字典存储关键词及其下的详细内容。接着读取之前抓取的评论，对每一个关键词看顾客的评论中是否有该词下属的详细词汇。这里用到"结巴分词"的分词功能，函数是 jieba.lcut。若找到了一个详细词汇，则将对应的单元格设为 1，反之设为 0。这里针对"结巴分词"的词库缺失，添加了部分词语。具体代码如下。

```
#!/usr/bin/python3
import xlwt
import xlrd
import jieba

# 要输出的表格
workbook = xlwt.Workbook()
sheet1 = workbook.add_sheet('sheet1',cell_overwrite_ok = True)

# 读取提取后的关键词
f = open('./classification.txt','r',encoding = 'UTF-8')
content = f.readlines()
f.close()

# 添加第一行的关键词名称
num = 0
```

```
writeNum = 0
keyWordTup = ()                                    # 所有的关键词
groupsDict = {}
while num < len(content):
    tem = content[num].find(' ')
    keyWord = content[num][0:tem]
    sheet1.write(0,writeNum,keyWord)
    keyWordTup += (keyWord,)
    temTup = ()
    groupContent = content[num].split(' ')
    for i in range(1,len(groupContent) - 1):
        temTup += (groupContent[i],)
    groupsDict[keyWord] = temTup
    writeNum += 1
    num += 1
# 读取抓取的数据
workbook1 = xlrd.open_workbook('./评论与评分.xlsx')
worksheets = workbook1.sheet_names()
worksheet1 = workbook1.sheet_by_name(u'othercommit')
customizedWords = ['海底捞','滑牛肉','海底捞牛肉','一根面','鸭血','柠檬水','抻面',
                   '海底捞笋片','午餐肉','豆花','鸭肠','支付宝','微信','牡丹园',
                   '地铁站','停车位','辣锅','鸳鸯锅','儿童','面筋','会员','免费',
                    '番茄锅','变脸','外送','小吃','虾滑','嫩牛肉','毛肚','鱼片','竹荪',
                   '猪脑','捞面','香蕉酥','手切羊肉','简阳鱼','小料','黄辣丁','油豆皮',
                   '宽粉','鱼豆腐','美甲','做指甲','免费水果','排号','棋牌','表演',
                   '锅底','半份','毛巾','哈密瓜','豆浆','ipad','零食','游乐场','果盘',
                   '车位','停车','游乐区','婴儿床','儿童乐园','外卖','外送','团购','下棋',
                   '贴膜','擦鞋','头绳','凉茶','果盘','糍粑']
for word in customizedWords:
    jieba.add_word(word)
num_rows = worksheet1.nrows
for curr_row in range(num_rows):                        # 对抓取数据进行遍历
    keyWordFlag = 0
    while keyWordFlag < len(keyWordTup):                # 对关键词进行遍历
        keyWord = keyWordTup[keyWordFlag]
        cell = worksheet1.cell_value(curr_row,4)        # 取评论
        cell_list = jieba.lcut(
            str(cell).lstrip().rstrip(),cut_all = True) # 分词
        find = 0
        for oneWord in groupsDict[keyWord]:             # 找关键词
            try:
                cell_list.index(oneWord)
                find = 1
                break
            except:
                pass
        if find == 1:
            sheet1.write(curr_row + 1,keyWordFlag,1)
        else:
            sheet1.write(curr_row + 1,keyWordFlag,0)
        keyWordFlag += 1
```

```
workbook.save('commentsWord.xls')
```

将最初抓取的数据中与评分相关的 4 列加入到代码生成的表格中，同样可以用 Python 实现，这里直接将之前的数据复制过来。comments Word. xls 如图 7.33 所示。

	A	B	C	D	E	F	G	H	I	J	K
1	评价均分	口味评分	服务评分	环境评分	底料	配料	面食	牛羊肉	其它肉类	河鲜与海鲜	豆制品
2	irr-star50	口味4(非常	服务4(非常	环境4(非常	1	1	0	0	0	0	0
3	irr-star50	口味4(非常	服务4(非常	环境4(非常	1	0	0	0	0	0	0
4	irr-star50	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
5	irr-star50	口味3(很好	服务4(非常	环境3(很好	0	0	0	0	0	0	0
6	irr-star50	口味4(非常	服务2(好)	环境2(好)	0	0	0	0	0	0	0
7	irr-star50	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
8	irr-star50	口味3(很好	服务4(非常	环境4(非常	1	0	0	0	0	0	0
9	irr-star50	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
10	irr-star40	口味4(非常	服务3(很好	环境3(很好	0	1	0	0	0	0	0
11	irr-star50	口味4(非常	服务4(非常	环境3(很好	0	1	0	0	0	0	0
12	irr-star50	口味4(非常	服务4(非常	环境3(很好	0	0	0	0	0	0	0
13	irr-star50	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
14	irr-star50	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
15	irr-star40	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
16	irr-star50	口味4(非常	服务4(非常	环境4(非常	0	0	0	0	0	0	0
17	irr-star40	口味3(很好	服务3(很好	环境3(很好	0	1	1	0	0	0	0
18	irr-star50	口味4(非常	服务4(非常	环境3(很好	0	0	0	0	0	0	0

图 7.33　comments Word. xls

将该表格导入 SPSS Modeler 18.0，单击“插入”→“源”→Excel，选择文件类型和导入文件。接着对评分的 3 个字段进行数值化处理。在之前的“源”节点后单击“插入”→“字段选项”，选择“导出”，将导出字段设置为“评价均分_数值”，并输入公式“substring_between(9,9,评价均分)”，单击“确定”按钮，如图 7.34 所示。

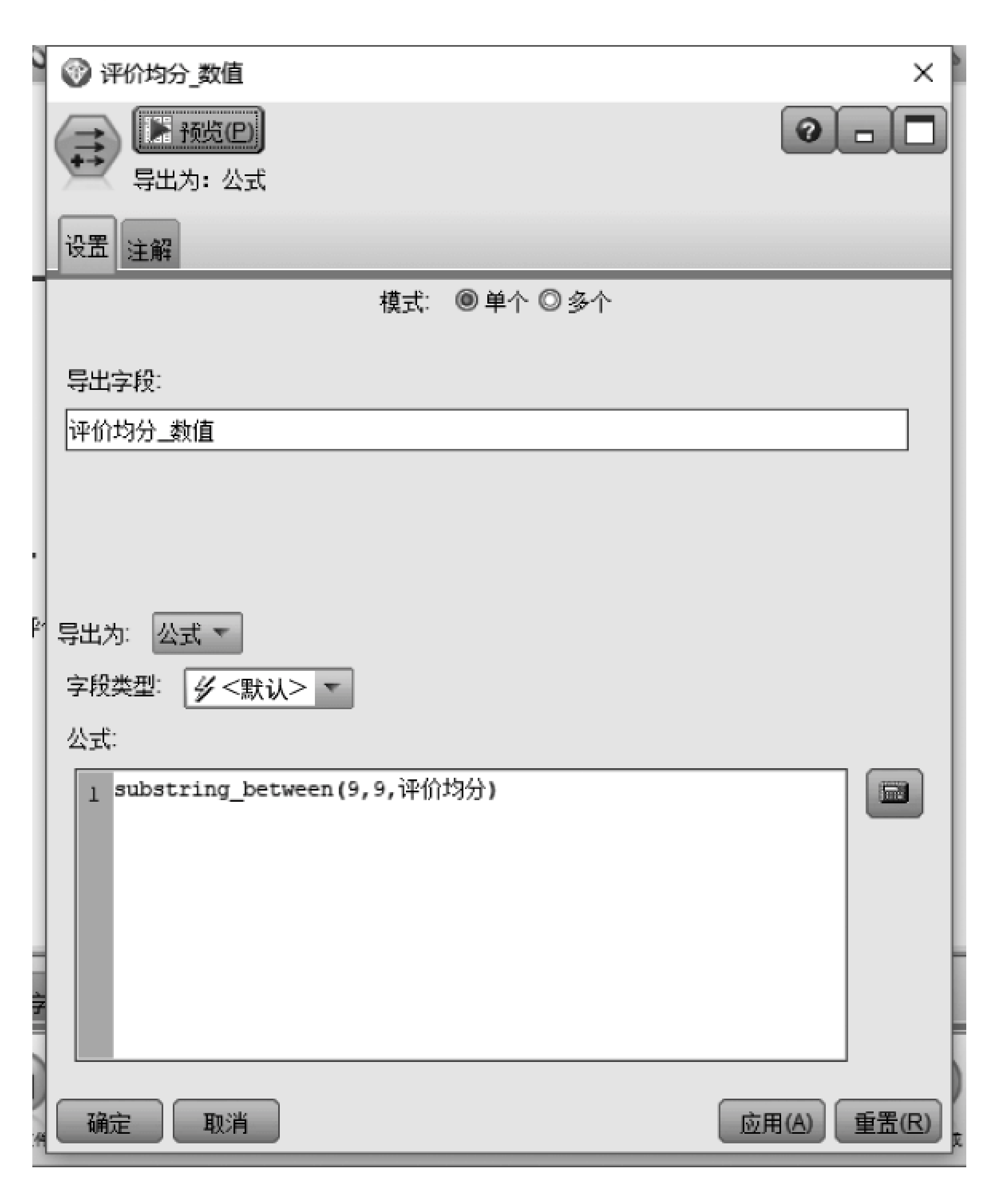


图 7.34　数值化“评价均分”

继续添加 3 个导出节点，分别对“口味评分”“服务评分”和“环境评分”做数值化处理。在 3 个节点后单击“插入”→“字段选项”，添加“过滤器”，将前 4 个字段过滤掉，单击“确定”按钮，如图 7.35 所示。



图 7.35 过滤数值化前的数据

可单击“插入”→“输出”，添加“表格”节点查看导出后的字段。

单击“插入”→“字段选项”，选择“类型”节点。将除评分外的其他字段的测量设为“标记”。将“评价均分_数值”的角色设为“目标”，将“口味评分_数值”“服务评分_数值”和“环境评分_数值”的角色设为“无”。其他保持默认值，单击“确定”按钮，如图 7.36 所示。



图 7.36 设置“类型”节点

在该“类型”节点后，单击“插入”→“建模”→Apriori，设置Apriori节点中的“最低条件支持度”为3.0，最下规则置信度为5.0，其他保持默认值，单击“运行”按钮得到结果模型。

这里，第一次生成的模型中很大比例都有“服务”，考虑到本店的服务一直是核心竞争力，为了更多地突出其他因素，在“类型”节点中将“服务”的角色设为“无”。重新运行生成模型。

选择按照“置信度百分比”排序，得到的结果如图7.37所示。可以对结果进行下列分析。这里的分析一部分是对本店进一步改进的建议，另一部分是其他火锅店可以学习本店的内容。

排序依据:置信度百分比　47 的 47

后项	前项	支持度百分比	置信度百分比
评价均分_数值 = 5	额外表演或服务	7.093	70.765
评价均分_数值 = 5	面食	5.988	66.019
评价均分_数值 = 5	半份	3.74	65.285
评价均分_数值 = 5	其他素菜 底料	4.593	64.979
评价均分_数值 = 5	回头客	7.888	64.373
评价均分_数值 = 5	其他素菜	7.791	63.682
评价均分_数值 = 5	朋友聚餐	13.004	63.189
评价均分_数值 = 5	河鲜与海鲜	8.469	63.158
评价均分_数值 = 5	零食	3.488	62.778
评价均分_数值 = 5	排队	21.395	62.5
评价均分_数值 = 5	其他肉类	6.822	61.932
评价均分_数值 = 5	朋友聚餐 排队	3.081	61.635
评价均分_数值 = 5	底料 排队	3.314	61.404
评价均分_数值 = 5	牛肉类	7.558	60.769
评价均分_数值 = 5	配料	8.062	60.577
评价均分_数值 = 5	底料	10.814	60.036
评价均分_数值 = 5	饮料	3.934	59.606
评价均分_数值 = 5	夜晚营业	3.062	54.43
评价均分_数值 = 4	朋友聚餐 排队	3.081	35.22
评价均分_数值 = 4	牛肉类	7.558	34.103
评价均分_数值 = 4	其他肉类	6.822	34.091
评价均分_数值 = 4	底料 排队	3.314	33.918
评价均分_数值 = 4	底料	10.814	33.692

图7.37　评论和打分关联分析

(1) 顾客在提到有额外表演或服务的时候，有高达70%的顾客给出了5分好评。其他店如果没有这方面的表演或服务，可以考虑增加这方面的内容，增加用户满意度。

(2) 在食材方面，评论涉及面食、河鲜与海鲜、牛肉类的，均有超过60%的顾客给予了好评，本店可以考虑结合上文做的菜品推荐和这里的高分评价菜，一起做菜品的营销。

(3) 同时提到“其他素菜”和“底料”的时候，5分好评占到64.9%。本店可以将“其他素菜”分类的素菜与“底料”一起组合成不同的锅底供顾客选择。

(4) 有高于8%的顾客提到配料，其中高于60%的顾客给予了5分好评，另外约有32%的顾客给予了4分好评，4分与5分评价总和超过90%。这说明本店的配料也有独到的地方，其他火锅店可以尝试学习改进自己的配料。

(5) 在提到“饮料”的评论中，5分评价达到59%，4分评论达到28%。而在分类中提到的饮料主要为“柠檬水”“豆浆”和“凉茶”3种，说明这3种饮料更受欢迎。其他火锅店也可以考虑增加缺失的饮料。

(6) 提到“半份”的有65%的顾客给予了好评,这也是本店的特色之一。菜品允许点半份,这既避免了浪费,也方便顾客点不同的菜。其他的火锅店可以考虑借鉴这个销售策略。

(7) 有21%的顾客提到“排队”,但其中62.5%的顾客仍然选择了好评。这是因为本店在顾客等位的时候提供了免费的零食和额外的服务,这在减少顾客流失和保证客源上有很大的帮助。有3.488%的顾客提到了专为排队提供的零食,他们中也有超过62%的顾客给予了5星好评。

(8) 有7.888%的顾客的评论中提到“回头客”相关的内容。这说明本店在口味和服务上有一致性,并且能够吸引顾客二次消费。

除此之外,还可以对本店的顾客评论做情感分析,了解顾客喜好。

7.6 顾客情感分析

为了对用户做出情感分析,需要获得用户在大众点评上对于该火锅店所做的评论的内容,抓取评论使用Python脚本。

由于网络平台中“水军”以及恶意评论等行为的存在,所以得到的评论内容有可能是不够好的数据,所以接下来需要先对得到的评论内容进行预处理,可以使用在线的去重工具去除重复内容,链接为http://quchuchongfu.renrensousuo.com。

只需要将要去重的文本复制到文本框中,单击“去重”即可达到去重效果。

要进行文本的中文分词处理,这里用一个Python的分词工具——结巴分词,只需要安装jieba包,就可以在Python中使用。

```
import jieba
import jieba.analyse
import jieba.posseg as pseg
```

之后,对评论的文本内容进行中文分词处理:核心代码如下:

```
with open("comment.txt",'rb') as wf,open("comemntsWord.txt",'w') as wf2:
    content = wf.read()
    freq_word = {}
    freq_flag = {}
    contents = pseg.cut(content)
    for word,flag in contents:
        if(len(word)>1):
            if(flag == 'c' or flag == 'cc' or flag == 'p' or flag == 't' or flag == 'r' or
flag == 'd'):
                pass
            else:
                #print word,flag
            if word in freq_word:
                freq_word[word] += 1
            else:
                freq_word[word] = 1

                #freq_flag[word] = temp
```

```
        freq_word_1 = []
        for word,freq in freq_word.items():
        freq_word_1.append((word,freq))
        freq_word_1.sort(key = lambda x: x[1],reverse = True)

        for word,freq in freq_word_1:
            if (freq > 10) :
            ws1.append([word,freq])
    wb.save(filename = dest_filename)
```

pseg.cut()方法针对文本进行分词，其中 word 和 flag 表示处理得到的关键词和该关键词的词性，由于处理得到的关键词中，介词、连词、时间词、代词、副词等是没有意义的，所以可以过滤除去这些词，同时计算关键词在该评论文本中的出现频率。最后将结果（关键词和该关键词的频率）保存到一个 Excel 文件中。

尽管通过过滤除去了介词、连词、时间词、代词、副词等，但仍然有一些词是没有意义的，这时可以手动去除没有意义的词。处理后的结果如图 7.38 所示。

服务	3895
不错	1711
味道	1296
好吃	964
喜欢	895
服务员	853
环境	633
排队	518
菜品	436
锅底	360
热情	321
吃饭	250
水果	231
免费	230
服务态度	229
美甲	227
价格	215
推荐	213
大家	207
贴心	204
没的说	191
时间	190
没得说	189

图 7.38　处理后结果

根据分词之后的词频，可以画出标签云，如图 7.39 所示。

从标签云图可以看出顾客的关注点主要在于“服务”“味道”“环境”等，也可以看出，这家火锅店能够为前来用餐的顾客提供“好吃”的菜品，让很多客户觉得“不错”，服务比较“热情”。这些优势是需要继续维持的方面。火锅店也可以针对这些特色做广告宣传的工作。但是，从标签云图同时也能够看到一些存在的问题，例如顾客会觉得店里过于拥挤，需要排队。火锅店可以考虑开设分店将消费者分流或者是制订避开高峰时间段用餐的优惠政策（折扣、礼品馈赠等方式），这是火锅店可以做出改善的细节。

下面再分析文本中用户的情感。这里用一个 Python 的情感分析包 SnowNLP 来实现，它会分析每条评论的用户情感，并给出一个[0～1]之间的数值，从 0 到 1 表示了消极情绪到积极情绪的变化过程。越靠近 1 说明积极情绪越高。处理的核心代码如下：

```
from snownlp import SnowNLP
import codecs
fr = open('comment.txt','r',encoding = 'utf - 8')
fw = open('motion1.txt','w',encoding = 'utf - 8')
while 1:
    line = fr.readline()
    if not line:
        break
    s1 = SnowNLP(line)
    fw.write(str(s1.sentiments) + " " + line)
```

s1.sentiments 得到该条评论的得分，并最终将每条评论的得分与该评论的内容写入到 motion1.txt 文件中，如图 7.40 所示。



图 7.39　标签云图

```
0.761632260522249      和老师在一起吃个饭，然后，服务好像变差了。
0.9897590820289026      一如既往地好，肉很好吃，小料不错。
0.5883256167466422      味道很好，小料种类超级多，水果也多，两人300不到
0.9821052018006514      服务员很热情，给樊娇娇点个赞!
0.9999997542655679      帅帅的服务生小哥很贴心的阻止了我盲目的点餐行为！超贴心！还会再去
0.9998453503119994      昨天晚上朋友过生日，专门定个六人的房间，何娇服务员服务非常棒，下回还会来！给你个赞赞赞\
(≧▽≦)/!

0.8473081734858704      大学七年来这里数不清多少次，服务没的说，味道也好，很受欢迎。
0.9999999253582337      来了很多次海底捞.服务态度非常好.美丽的小燕服务很到位很贴心.微笑服务.真的很棒.谢谢燕子的服务.
一起为美好的明天加油!
```

图 7.40　评论与情感得分

可以看出，得分基本反映了用户的情感，是比较合理的。

那么，根据该数值，就可以得到用户的情感。划定积极情绪、中间情绪以及消极情绪之间的范围为[0.6～1]、[0.4～0.6]、[0～0.4]。然后计算积极情绪、中间情绪和消极情绪的比例。核心代码如下：

```
while 1:
    line = f.readline()
    if not line:
        break
    t = line.split(' ')
    w = t[0]
    m = t[len(t) - 1]
    if float(w)> 0.6:
        count_good = count_good + 1
        q.write(m)
    elif float(w)> 0.4 and float(w) <=  0.6:
        count_temp = count_temp + 1
        tc.write(m)
    else:
        p.write(m)
        count_bad  +=  1
    count = count + 1
    temp = line
```

通过执行代码，得到各种情绪占比。

积极情绪、中间情绪、消极情绪的比例分别为 0.77、0.05、0.17，大部分客户对该店是比较满意的。态度处于中间水平的客户不多，有将近两成的客户对该店不满意。

可以从热评词中分析关注该热评词的客户对该店的态度：从上面的标签云图中很容易发现评论的热评词，如“服务”“味道”“环境”等，可以从这些热评词入手，看客户比较关注的地方，该店做得怎么样。从最热的词“服务”入手，计算出提到“服务”这一热评词的所有评论有多少，然后分析其中的积极情绪、中间情绪、消极情绪分别占多少比例。

```
for x in range(1,len(all_food) + 1):
    comment = codecs.open('goodCom1.txt','r','utf - 8')
    comment_temp = codecs.open('tempCom1.txt','r','utf - 8')
    comment1 = codecs.open('badCom1.txt','r','utf - 8')
    cou_good = 0
    cou_temp = 0
    cou_bad = 0
    while 1:
        line = comment.readline()
        if not line:
            break
        if line.find(all_food[x - 1])!= - 1:
            cou_good = cou_good + 1
    while 1:
        line = comment_temp.readline()
        if not line:
            break
        if line.find(all_food[x - 1])!= - 1:
            cou_temp = cou_temp + 1
    while 1:
        line = comment1.readline()
        if not line:
            break
        if line.find(all_food[x - 1])!= - 1:
            cou_bad = cou_bad + 1
    print str(cou_good),str(cou_temp),str(cou_bad)
    filee.write(all_food[x - 1] + " " + str(cou_good) + " " + str(cou_temp) + " " + str(cou_bad) + "\n")
```

执行之后，就得到相应的数据，如图 7.41 所示。

为了对比分析，可以找出评论中不包含“服务”这一热评词的评论，分析其各种情绪所占的比例，结果如图 7.42 所示。

服务
积极情绪: 0.7980613893376414
中间情绪: 0.04819601507808293
消极情绪: 0.15374259558427572
总评论数: 3714

图 7.41　包含“服务”一词情绪分析

服务
积极情绪: 0.7014512785072564
中间情绪: 0.07049067035245335
消极情绪: 0.22805805114029026
总评论数: 1447

图 7.42　不含“服务”一词情绪分析

对比图 7.41 和图 7.42 可以发现，在不包含“服务”的评论中，积极情绪降低了将近 10%，消极情绪和中间情绪都有所增加，这说明“服务”这一因素很大程度上决定了该店的客户情感，所以说明“服务”是影响客户情感的关键因素，必须引起商家的重视。从包含“服务”这一热评词的情绪与所有评论的情绪的对比中可以看出，包含“服务”的评论中，积极情绪要比总体的情绪稍高，这说明该店在服务方面做得比较好，有一定的竞争力。所以，商家必须重视“服务”在客户情感中的重要作用，应该在“服务”方面继续保持优势，争取服务水平更上一层楼。

第 8 章

商务宾馆竞争分析

随着经济的发展和人们生活水平的提高，国内经济型酒店发展迅速，近几年有大量的经济型商务酒店建成并投入使用，竞争异常激烈。如何能在酒店行业的红海市场中生存和发展，是目前酒店的经营者一直在思考的问题。

通过对点评网站中的酒店评价数据进行抓取，获得用户对酒店的评分和评论内容，结合评论人、评价数量、评价内容、评价频次及评分随时间变化的走势，可以对酒店中存在的主要问题进行分析，并对用户复购率进行统计，结合线性回归等数据分析算法统计分析得到酒店竞争力影响因素。对评价内容进行词频统计和情感分析，综合正面、中立、负面情感对酒店的竞争情况进行比较，最终得出酒店的竞争过程中存在的主要问题，并给出改进建议，从而提高酒店的市场竞争能力。

8.1 目前经济型酒店行业竞争态势

2003 年之后，随着如家、7 天、汉庭等本土经济型酒店品牌的创建，本土经济型酒店迅猛发展。根据盈蝶咨询数据统计，至 2015 年 1 月 1 日，本土经济型酒店的门店总数已经达到 15 439 家，客房数共 1 525 471 间，品牌数共 514 个。根据 2014 年上市公司财报，本土经济型酒店品牌按市场占有率排名，前 10 名依次是：如家快捷、7 天酒店、汉庭酒店、锦江之星、格林豪泰、莫泰、玖玖旅馆、尚客优、布丁酒店、城市便捷。

经济型酒店因为服务、环境等标准化，并且相对来说同一品牌的质量相对稳定，最重要的是具有高性价比，所以经济型酒店发展迅速，几年时间引来大量资本投入，导致经济型酒店进入红海时代，同质化竞争严重，利润下降，使得服务水平降低，影响用户体验，最终整个行业进入低赢利水平阶段。

开一家经济型酒店的成本并不高，进入门槛较低，容易受到新进入者的威胁；随着人们生活水平的提高，人们对星级酒店消费能力的提升，对经济型酒店产生空间挤压，即“白领用户”趋于选择星级酒店，而“蓝领用户”对价格较敏感，对偏高端的经济型酒店具有排斥心态，更愿选择更低价格的简陋旅馆。由于客人的选择范围广，所以他们会议价，压缩酒店的利润空间。综合来看，目前酒店行业竞争态势还处于低层次竞争级别，差异化不明显，更多的是进行价格竞争，要想在众多酒店中获得更大的优势，主要还需要对服务进行创新，通过服务好目标客户群产生良好的品牌口碑，逐渐在竞争中胜出。

如图 8.1 所示，A 商务宾馆是一家定位为较高性价比的经济型商务酒店，其位于高铁站附近，直线距离高铁东站不超过 500m。酒店拥有大床房、温馨家庭房、舒适三人间、舒适双床房、舒适大床房、观景双大床房、阳光双床房、阳光商务大床房等房型，房内配设齐全，提供 24 小时热水、空调、卫浴、电视、电话等配套设施。

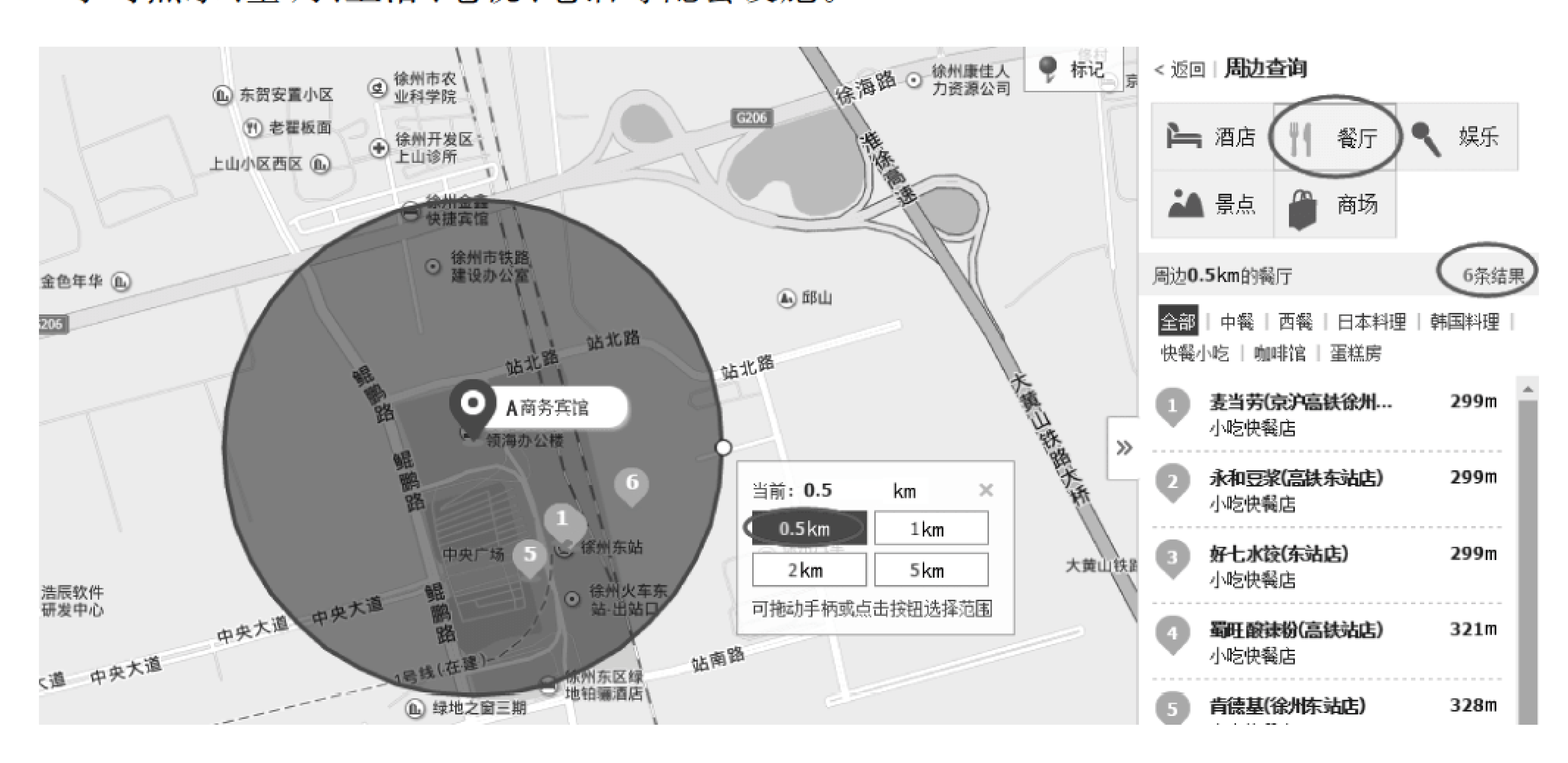


图 8.1 商务宾馆附近餐饮情况展示

服务、位置、餐饮、配套等是酒店的重要影响因素，A 商务宾馆距离火车站非常近，直线距离只有 0.5km，并且在商务宾馆 2km 内有两个商业广场，具有较高的客源流量基础，这是其重要的竞争优势。

虽然 A 商务宾馆占据了交通枢纽的优势，但其面临的威胁是酒店行业进入门槛较低，吸引了众多的行业竞争者进入，在酒店附近存在数十家同样价位的商务宾馆，由于差异化不明显，竞争者都采用针对式的营销策略，在开发新客户时推出层出不穷的优惠方案，对 A 商务宾馆的经营产生较大影响。另外，虽然交通枢纽处的餐饮较多，但其质量和价格参差不齐，对 A 宾馆易带来负面评价。综上，A 商务宾馆要想在竞争中立于不败之地，需要具有较强的竞争能力和客户服务能力。

能否在竞争中脱颖而出取决于如何进行差异化经营，如何增强商务宾馆的商业竞争力，为了实现这一目标，通过对入住客人的评论内容进行不同维度的数据挖掘，获取 A 商务宾馆竞争现状和存在的主要问题，并为其提出酒店经营的合理化建议，以提高其市场竞争能力。

8.2 用户相关数据准备

为了分析 A 商务宾馆目前在其行业内的竞争情况、客户满意情况以及客户反馈的主要问题，需要获得顾客对酒店的评论内容和评价分数，这些数据可以通过使用“爬虫”软件或编程的方式从酒店预订网站上抓取。不仅可以抓取 A 商务宾馆的客户评论数据，还可以抓取其周围竞争对手酒店的评论数据，作为对比分析依据。结合网站上的客户点评数据，可以提取客户对酒店的评分、评论内容、评论人、评价数量、评价人等级等信息，并将上述文本内容进行格式化存储，用于后续的数据分析。

1. 使用软件工具抓取评论

由于网站上的评论数据成千上万条，靠人工整理效率低下且易出错，可以使用“八爪鱼(http://www.bazhuayu.com)”等工具软件实现内容自动抓取，操作过程简单快速，其原理是模拟浏览器对网站的浏览，在页面加载完成之后，通过提取页面 HTML 代码中的对应节点的文本内容来获取网站上的目标数据，数据提取之后以文本文件方式存储，对于海量评论数据，还可提供云端抓取和下载。

2. 用户评论内容抓取

“携程”网站上对酒店的评论页面是分页显示评论内容的，需要制作两级规则来抓取数据：第一级规则模拟单击“更多点评”按钮，并设置爬虫路线链接到下级规则；第二级规则通过重复单击下一页的爬虫路线抓取数据。

1) 抓取规则设置

首先命名主题名为“A 商务宾馆 demo”，规则编号默认，页面地址就是需要抓取数据的网址，如图 8.2 所示。注意，要对主题名进行查重，确定没有使用过即可进入下一步。



图 8.2 数据抓取命名

然后在“创建规则”的整理箱操作区单击“新建”，命名为“列表”。单击浏览器中的“查看全部点评”，定位到该按钮的 div 节点，如图 8.3 所示。

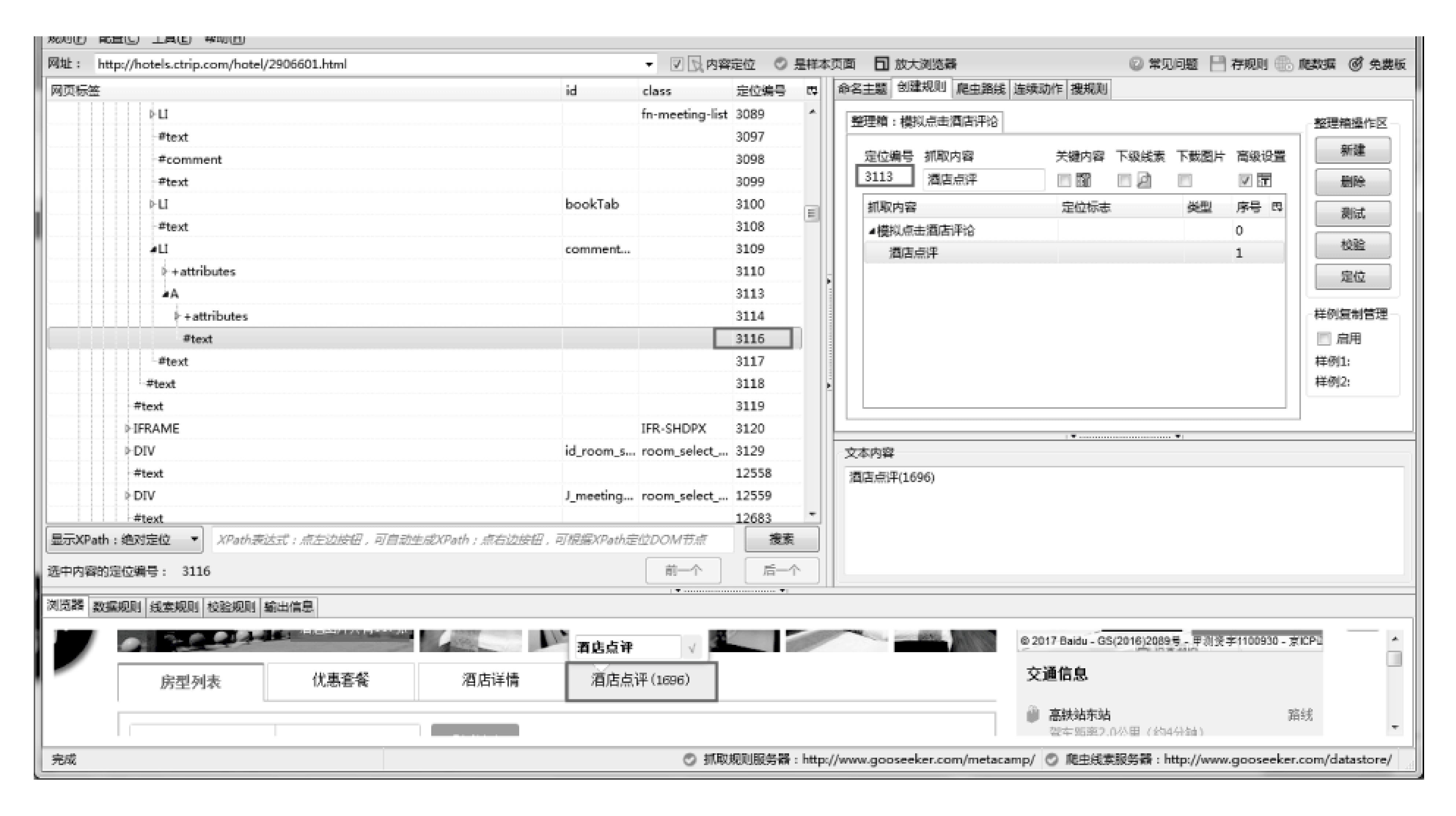


图 8.3 创建规则-新建

然后如图 8.4 所示，右击“＃text”，选择“内容映射”，将文本映射到设置的抓取内容“模拟单击标志”。



图 8.4 创建规则-选择映射

上述操作是模拟人在浏览网页过程中单击了“查看全部点评”，使用页面跳转到酒店评论的列表页面，供后续循环单条抓取。

2）设置爬虫路线

步骤如下：在操作区单击“新建”新建线索，并命名为“线索一”；设置“连贯抓取”，选择记号线索；单击“查看更多点评”定位到其对应的 div 节点，展开节点下的文件夹，找到 text

节点。右击“text 节点”，选择“线索映射”→“定位”→“线索一”。然后选择“线索映射”→“记号映射”，如图 8.5 所示。设置目标主题名，这是很关键的一步，因为这是接下来的主要抓取规则。这里的目标主题还不存在，接下来就要去建立这个抓取规则。

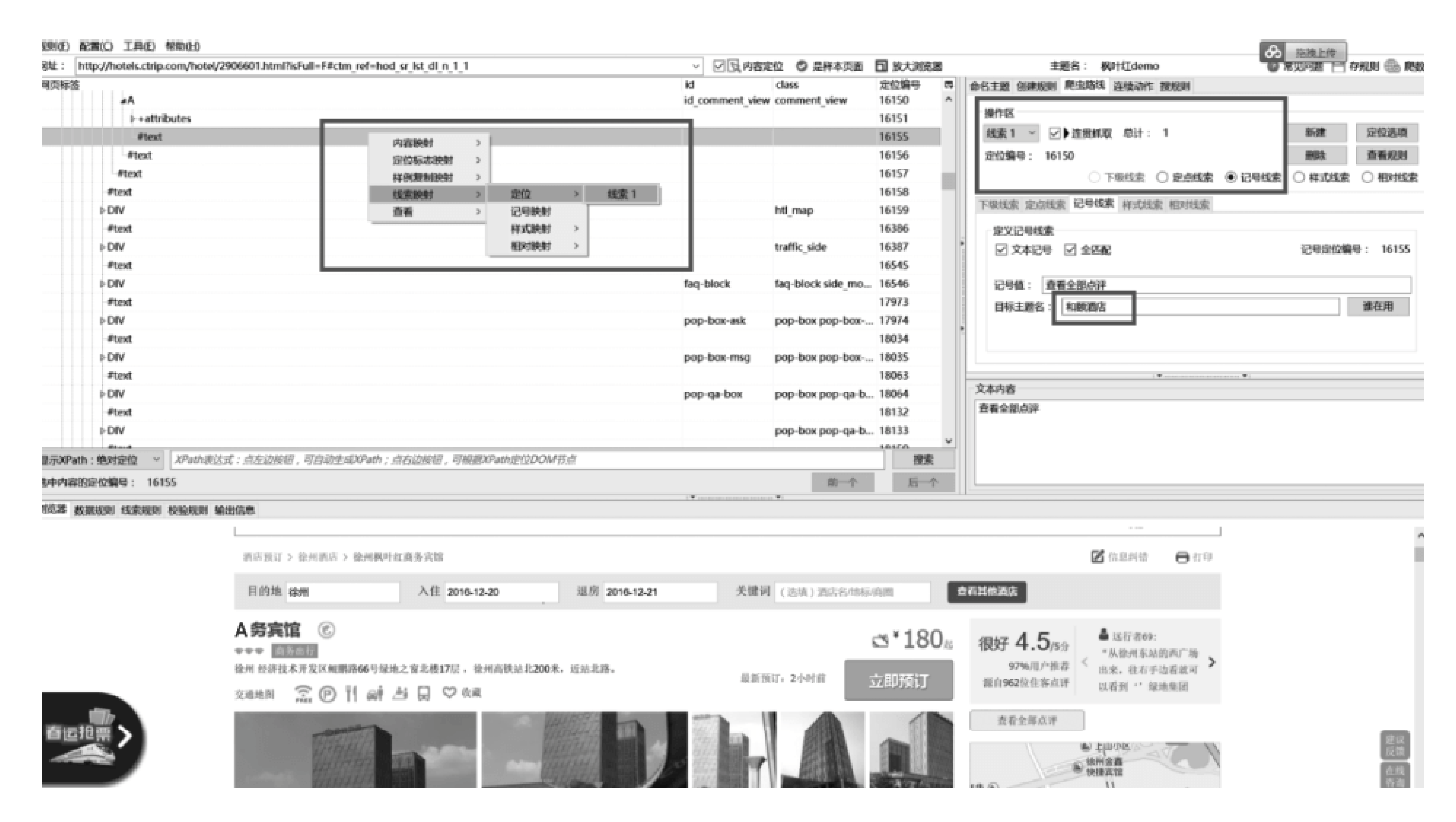


图 8.5 设置爬虫路线

爬虫路线设置完成表示整个列表的循环抓取开始，软件将按照设置的路线逐个抓取单条评论的所有相关内容。

3）第二级抓取规则

首先命名主题名为在第一级规则里设置的目标主题名，即“B 酒店”。需要抓取的内容主要是在用户评论中包含的内容。图 8.6 主要包括与用户有关的信息：用户昵称，用户等级，用户历史评论情况（点评总数，评论被点有用次数，上传图片总数）；用户的订房信息：出游目的，入住时间，入住房型；用户的评论信息：评论内容，评分（包含分类评分），评论发表时间。



图 8.6 抓取内容

新建整理箱，添加需要抓取的内容，针对不同的抓取内容需要不同的抓取方法，下面主要介绍几种特殊的抓取规则。

1）用户等级抓取

因为用户等级的 div 节点不是 text，所以没办法直接抓取文本。只能通过高级设置的自定义 xpath 来抓取 class 的文本。步骤如下：单击“点评新星”，找到对应的 div 节点，展开节点找到如图 8.7 所示的“@class”节点；在该节点处右击，选择内容映射到“用户等级”；然后选择抓取内容中的“用户等级”，单击“高级设置”；选择“自定义 xpath”“文本内容”，设置抓取内容表达式如图 8.7 所示，并保存。

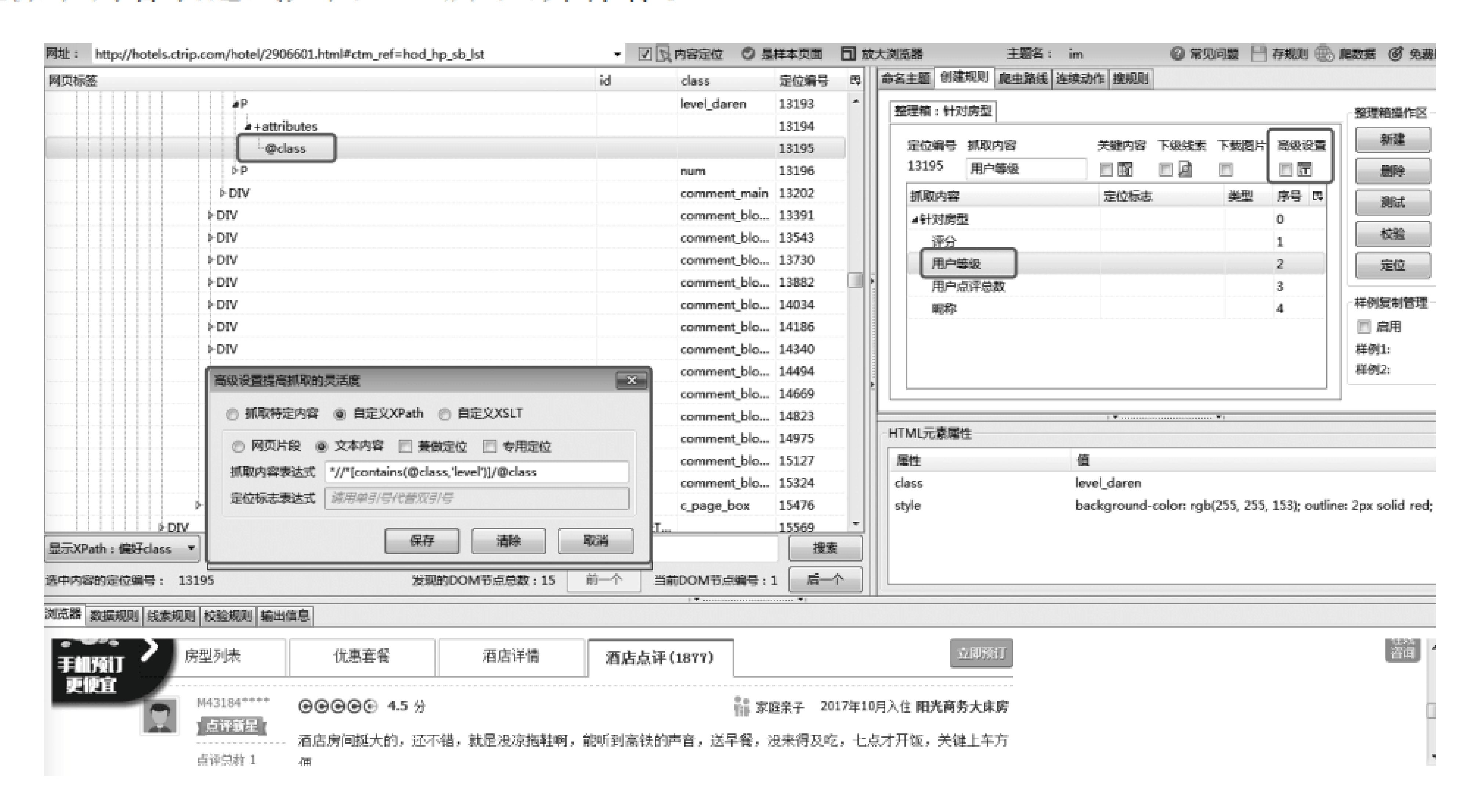


图 8.7 用户等级抓取

2）点评总数、评论被点有用次数、上传图片总数抓取

因为只有鼠标移动到“评分”区域时，这些信息才会悬浮显示，所以不能直接抓取到 text 信息。找到用户头像对应的节点，在节点对应的属性找到这些数据存放的节点，如图 8.8 所示的@data-usefulcount 和@data-img-count 等，右击“内容映射”即可。

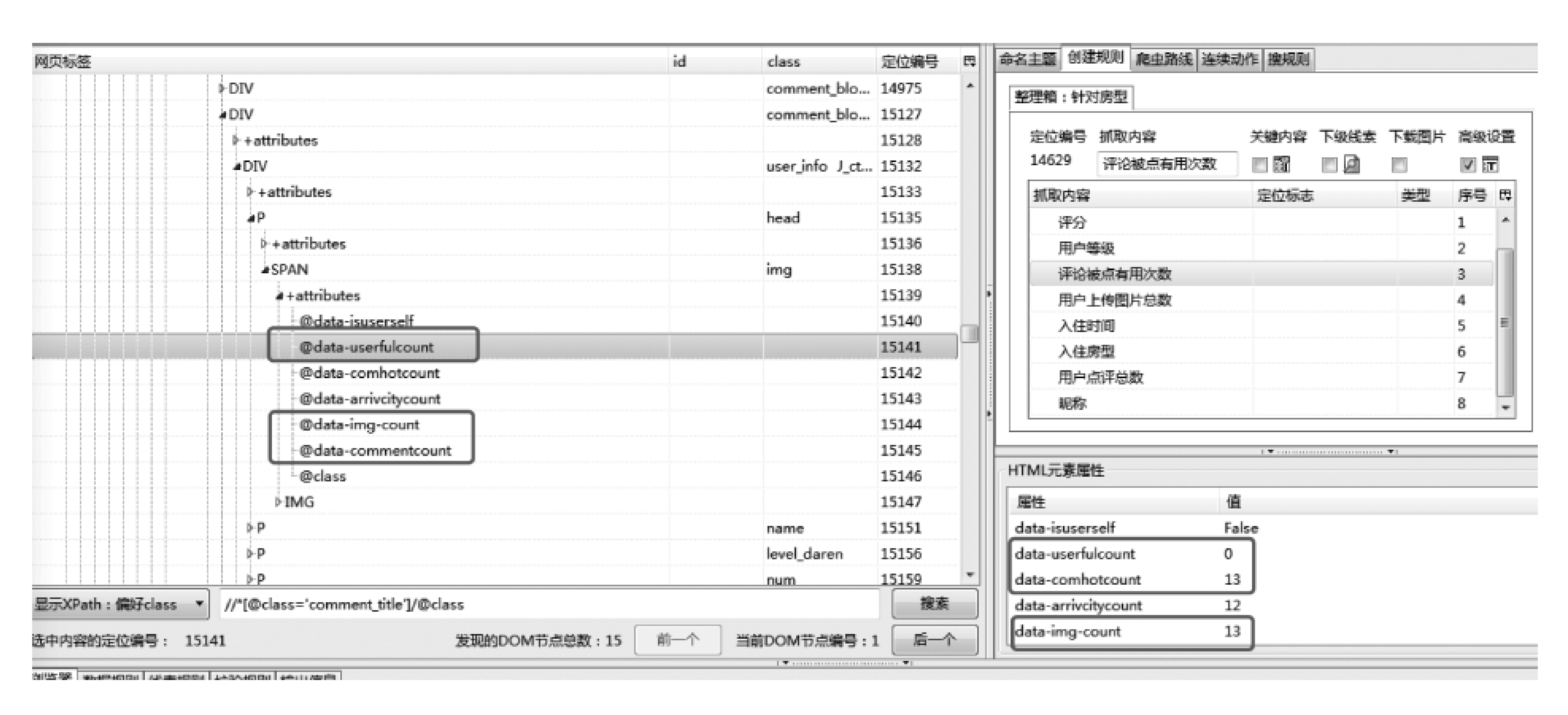


图 8.8 点评总数、评论被点有用次数、上传图片总数抓取

3）入住房型抓取

虽然房型对应的节点有 text 文本，但通过查看网页，发现并不是所有的房型节点都有类似的 text 节点，会出现漏抓的情况，所以通过自定义 xpath 的方式抓取比较稳妥。步骤如下：单击“和颐高级大床房”，找到对应的 div 节点；找到属性里的“@data-baseroomname”，右键单击它并选择内容映射到入住房型；单击“入住房型”，选择“高级设置”；在弹出的界面里选择“自定义 xpath”和“文本内容”，设置抓取内容表达式如图 8.9 所示，最后保存。

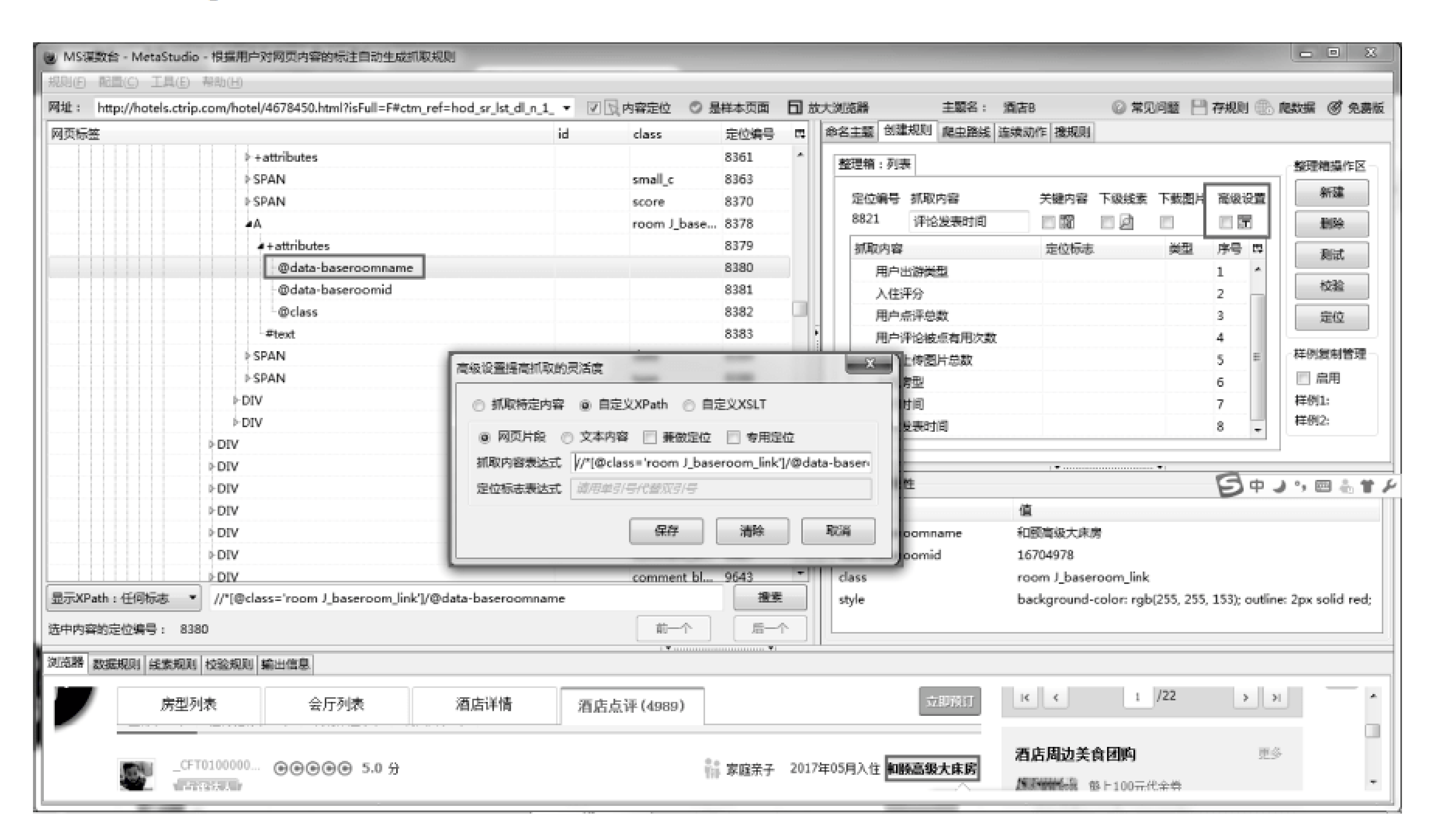


图 8.9 入住房型抓取

4）各类评分抓取

对分类评分信息的抓取是通过评分节点的@data-value 属性，如位置、设施、服务、卫生评分。步骤如下：单击图 8.10 所示的评分区域，找到相应的 div 节点。展开节点，找到包含“@data-value”的节点，右击“内容映射”即可。



图 8.10 各类评分抓取

如图 8.11 所示，单击“下一页”，找到“下一页”对应的节点，并设置内容映射，目标主题名依然是当前规则。这样就可以得到对每一页重复抓取的路线。

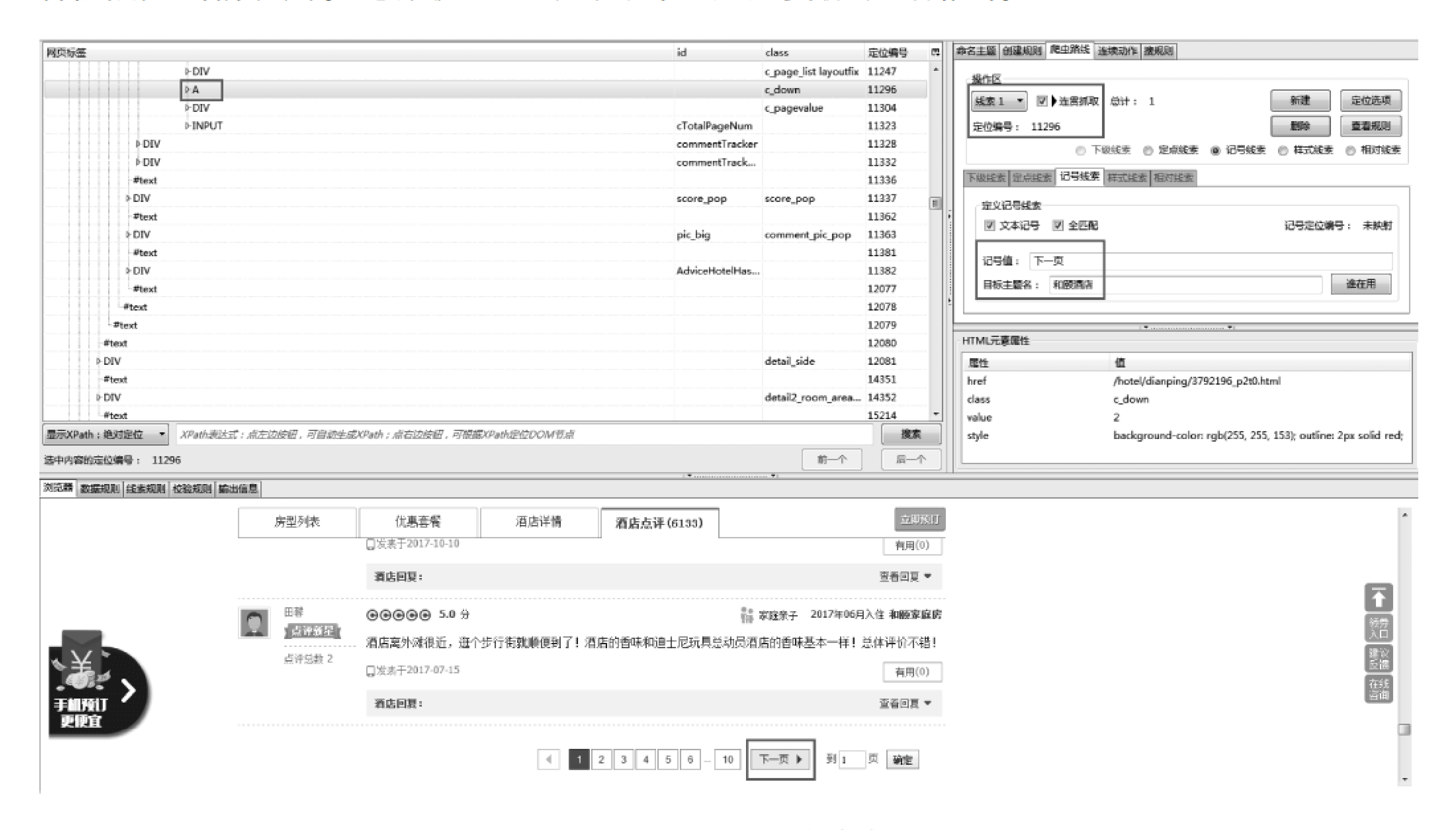


图 8.11　设置爬虫路线

在 DS 打数机里添加线索，即需要抓取内容的网站，单击“单搜”按钮即可。

图 8.12　数据抓取

通过 DS 打数机的抓取，得到一系列 xml 文件，每一页内容对应一个。通过“八爪鱼”里的转换功能将这些文件打包转换为 Excel，如图 8.13 所示。

将 xml 文件生成压缩包，然后使用“导入数据”，刷新后即可导出数据，如图 8.14 所示。

枫叶红demo1_278255077_299374619.xml	2016/12/18 20:11	XML 文档	10 KB
枫叶红demo1_278255077_299383089.xml	2016/12/18 20:11	XML 文档	10 KB
枫叶红demo1_278255077_299395680.xml	2016/12/18 20:11	XML 文档	11 KB
枫叶红demo1_278255077_299404145.xml	2016/12/18 20:12	XML 文档	10 KB
枫叶红demo1_278255077_299412621.xml	2016/12/18 20:12	XML 文档	10 KB
枫叶红demo1_278255077_299421113.xml	2016/12/18 20:12	XML 文档	9 KB
枫叶红demo1_278255077_299429600.xml	2016/12/18 20:12	XML 文档	8 KB
枫叶红demo1_278255077_299438084.xml	2016/12/18 20:12	XML 文档	9 KB
枫叶红demo1_278255077_299446484.xml	2016/12/18 20:12	XML 文档	9 KB
枫叶红demo1_278255077_299454830.xml	2016/12/18 20:12	XML 文档	9 KB
枫叶红demo1_278255077_299463272.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299471740.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299480216.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299488723.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299497225.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299505744.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299514204.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299522675.xml	2016/12/18 20:13	XML 文档	9 KB
枫叶红demo1_278255077_299531207.xml	2016/12/18 20:14	XML 文档	9 KB
枫叶红demo1_278255077_299539690.xml	2016/12/18 20:14	XML 文档	9 KB
枫叶红demo1_278255077_299548166.xml	2016/12/18 20:14	XML 文档	9 KB
枫叶红demo1_278255077_299556623.xml	2016/12/18 20:14	XML 文档	9 KB
枫叶红demo1_278255077_299565131.xml	2016/12/18 20:14	XML 文档	9 KB
枫叶红demo1_278255077_299573684.xml	2016/12/18 20:14	XML 文档	9 KB
枫叶红demo1_278255077_299582158.xml	2016/12/18 20:14	XML 文档	8 KB

图 8.13　数据抓取结果——XML 展示

图 8.14　数据抓取结果——Excel 展示

8.3　通过 Python 编程抓取评论

由于“八爪鱼”软件为收费软件，未付费用户仅能获取少量数据，并且其原理是模拟浏览器的浏览过程，需要等整个网页都加载完成后才可以提取内容，网站上的广告、图片等显示耗费较多网络资源，抓取时间耗时较长。为了快速提取评论内容，使用 Python 编程的方式，调用 Firefox 浏览器插件功能，实现从网络上爬取商务宾馆的评价信息和评分，其优点除了抓取时间短、效率高之外，更重要的是可以通过编程的方式实现抓取内容和过程的定制化。

这里主要基于 Python 语言实现，因此所用爬虫框架及网站解析框架均使用 Python 第三方包。

（1）下载并安装 Python 2.7 版本，下载网址为 https://www.python.org/downloads/。

（2）下载并安装 getckodriver，这是操作火狐浏览器的驱动，下载网站为 https://github.com/mozilla/geckodriver/releases。

（3）安装火狐浏览器。

（4）下载并安装 scrapy，这是 Python 的爬虫框架，可通过命令 pip install scrapy 安装，也可到 scrapy 网站直接下载包安装。

（5）下载并安装 selenium，安装命令为 pip install selenium。

（6）实验前需要实验者具备一定的 Python 基础知识，并且需要了解 scrapy 的基本使用，需要熟悉 xpath 语法。安装好 selenium 后需要参考它的说明文档了解 webdriver 的使用。

编写 Python 脚本的过程如下：

（1）创建项目。

安装好 scrapy 后，开启 cmd 窗口，可通过命令 scrapy startproject hotel 来创建 hotel 项目。如果系统无法识别 scrapy，则把 Python 的环境变量路径加到系统中。默认 Python 安装到 C 盘，则需要添加的环境变量为 C:\\Python\\和 C:\\Python27\\Scripts。

项目创建成功后会产生一个 hotel 文件夹，里面包含 hotel 文件夹和 scrapy.cfg 文件。进到 hotel 文件夹下的 spiders 文件夹，创建 hotel_spider.py 文件。在 hotel 文件夹下的 settings.py 文件中进行如下配置：

```
BOT_NAME = 'hotel'
SPIDER_MODULES = ['hotel.spiders']
NEWSPIDER_MODULE = 'hotel.spiders'
```

至此，项目创建完毕，接下来主要修改 spiders 文件夹下的 hotel_spider.py。

（2）编写规则。

修改 hotel_spider.py 文件，增加 HotelSpider 类，该类继承于 Spider 文件，需要导入如下包：

```
from selenium import webdriver
from scrapy.spiders import Spider
from scrapy.selector import Selector
from scrapy.http import Request
import requests
```

为 HotelSpider 类添加类变量 name，allowed_domains，start_urls。其中，name 为该爬虫的名字，allowed_domains 为允许爬虫爬取的网站列表，start_urls 为爬虫开始爬取的网站列表。以爬取携程网为例，这 3 个参数可以设置如下（2906601.html 为 A 商务宾馆在携程的网页）。

```
name = "hotel" allowed_domains = [r"http://hotels.ctrip.com/hotel/2906601.html"]
start_urls = [r"http://hotels.ctrip.com/hotel/2906601.html",]
```

接下来编写 parse 函数，当爬虫开始工作后，parse 函数将承担解析工作。

parse 函数为被动触发函数，参数为 response，这个 response 包含所有网页和请求返回

的所有信息。response 支持 xpath 解析，xpath 为爬虫的主要解析方式。

用火狐浏览器打开 A 商务宾馆在携程的网页，并右键选择查看元素，则出现火狐的网页调试窗口，如图 8.15 所示。



图 8.15 元素提取页面

点选调试窗口最左侧的元素选择图标，再单击页面上需要定位的元素，则调试器窗口内会显示该元素的 html 源码。如果需要解析这部分内容，则需要编写相应的规则。例如，需要提取标题，则先找到标题的 html 元素：

```
<h2 class="cn_n" itemprop="name">A 商务宾馆</h2>
```

通过 xpath 解析标题，写法：

```
'//h2[contains(@class,"cn_n")]/text()'
```

使用 response 提取的完整写法为：

```
response.xpath('//h2[contains(@class,"cn_n")]/text()').extract()
```

这样可提取 h2 标签，并且 class 为 cn_n 的元素，即我们需要的标题。

网页上的其他内容信息原理相同，当 xpath 匹配多条时，则返回列表，这在解析客户评价的时候非常有用。

通过解析 http://hotels.ctrip.com/hotel/dianping/2906601.html 可获得点评数据，因为点评数据为分页数据，所以通过 Python 的循环爬取，把网页链接存到列表里，然后依次解析并保存每一个网页的评价。

可以通过 selenium 的 webvriver 模块操作火狐浏览器来解析每一个评价网页。

```
browser = webdriver.Firefox()
browser.get(url)
```

以上两行为获取一个火狐浏览器并加载 url 到火狐浏览器内。完成后，browser 有方法可进行 xpath 解析。

某些网站可能通过 post 请求来获取数据，此时可使用 Python 的 requests 模块来实现

post 的交互。

通过命令行 scrapy crawl hotel 即可激活爬虫爬取数据，经过 parse 函数解析后保存到文件。

本实验的 demo 源码 parse 内实现的是对艺龙网的爬取，如需爬取携程，则可将 parse 函数注释，修改 parse_ctrip 函数名为 parse，并且修改 start_urls 内的网站为 ctrip 的网站。爬取的网站数据分别会保存到 log. txt 以及 elong. txt 中。

8.4 数据预处理

抓取到的数据非标准化内容较多，特别是用户评价具有较大的随意性，直接应用会对分析结果产生较大干扰，所以需要对数据进行预处理，包括异常数据过滤和数据整理。

首先将数据库随意性评论数据进行删除，如一长串的“好”“不错”等字，随意输入的英文字母等，剔除 26 条。对用户评分空值或者明显异常的数据进行筛选去除。在获取的原始数据集中，异常值均为空值，将为空值的评价数据直接剔除。由于酒店评价是按照房间进行的，如果某客人一次订了多间房，可能会重复评论相同的内容，这部分数据易影响词频分析，故将同一人同一次入住的重复评论剔除。将某些凑字数的评论中重复输入的文字移除，只保留其中一个。如经初步筛选及过滤的结果如图 8.16 所示。

入住时间	评论发表时间	评论	位置	设施	服务	卫生
016年10	2016/10/8	从徐州东站的西广场出来，往右手边看就可以看到	4	4	4	4
016年11	2016/11/6	酒店上次入住就没找到地方，跟着导航挺难找的。	3	4	5	4
016年10	2016/10/4	设施还不错，挺干净，就是晚上走廊脚步声太多！	5	4	5	5
016年11	2016/11/27	有点尴尬一次入住，客房紧张没有靠边的房间了，	5	5	5	5
016年10	2016/10/7	酒店新开，位于高铁站斜对面的右手边的绿地集团	5	5	5	5
016年11	2016/11/17	高铁东站旁边，非常方便。酒店前台很热情，酒店	3	5	5	5
016年10	2016/10/18	房间非常干净整洁，宽敞明亮，服务也很好，离高	5	5	5	5
016年12	2016/12/9	不错吹吹噢，不错up的，不错的啊。。。。。。离	5	5	5	4
016年09	2016/9/11	过来这边就选这家酒店，离高铁比较近，挺好的，	5	5	5	5
016年06	2016/7/29	离火车站很近， 该酒店下次去还会住的，交通很方	4	4	4	4
016年11	2016/11/23	酒店离高铁站很近，步行五分钟。唯一不足就是指	5	5	5	5
016年11	2016/11/13	性价比非常高的一个酒店。下楼斜对面就是高铁站	5	5	5	5
016年08	2016-08-18(本	干净的酒店，出门就是高铁。晚上零点楼层异响持	5	5	3	5
016年11	2016/11/9	房间不错很安静，美中不足就是周边吃饭的地方太	4	4	4	4
016年10	2016/10/14	离徐州东站非常近，走几分钟就到。室内打扫挺干	4	4	4	4
016年10	2016/10/5	非常好的商务酒店，绿地开发的房子第一印象就不	5	5	5	5
016年10	2016/10/7	在徐州东站转车住了一晚。酒店大堂服务态度好，	4	5	5	5
016年12	2016/12/14	离高铁站近，方便	4	4	4	5

图 8.16 经初步筛选及过滤的结果

情感分析、主题分析的对象都是用户的评论数据，需要对数据进行必要筛选，以提高后续分析的质量，其中 A 商务宾馆的评论数据条数为 751 条，B 酒店的为 2351 条，C 商务宾馆的为 1050 条。将抓取到的 txt 格式文本复制到空 Excel 文件中形成格式化文档，新建数据库表，表结构与 Excel 中的列名一致，另外新增一个酒店名称字段，用于区分酒店的评论，将 3 家酒店的数据分别导入，形成统一的数据集合。

为了更好地分析评分走势，新增字段累计综合评分和累计评论总数，以周为单位统计平均分和评论数，将平均值按照周的顺序逐周累加，形成累计平均综合评分和累计评论总数。

8.5 商务宾馆客户数据分析

在数据获取和预处理之后，通过对 A 商务宾馆的评论数据进行建模分析，得到宾馆在各个方面的点评结果。首先分析宾馆评分的主要影响因素，并依此对酒店进行基础分析、消费者决策分析、与同类酒店的竞争分析。

8.5.1 酒店评分影响因素

数据整理之后依然具有较多维度，通过使用决策权分析、Apriori 关联分析对酒店评价的主要影响因素进行分析，获取与评分高低相关的变量。

将 A 商务宾馆、B 酒店、C 商务宾馆的评论数据中的综合评分以箱图的格式显示，如图 8.17 所示，可以看到评分主要分布为 4.0～5.0 分，说明 B 酒店的分值最集中，几乎接近 5.0 分，而 A 商务宾馆次之，综合评分在 4.2 以上，C 商务宾馆的评分分布较差，分值为 4.0～5.0。

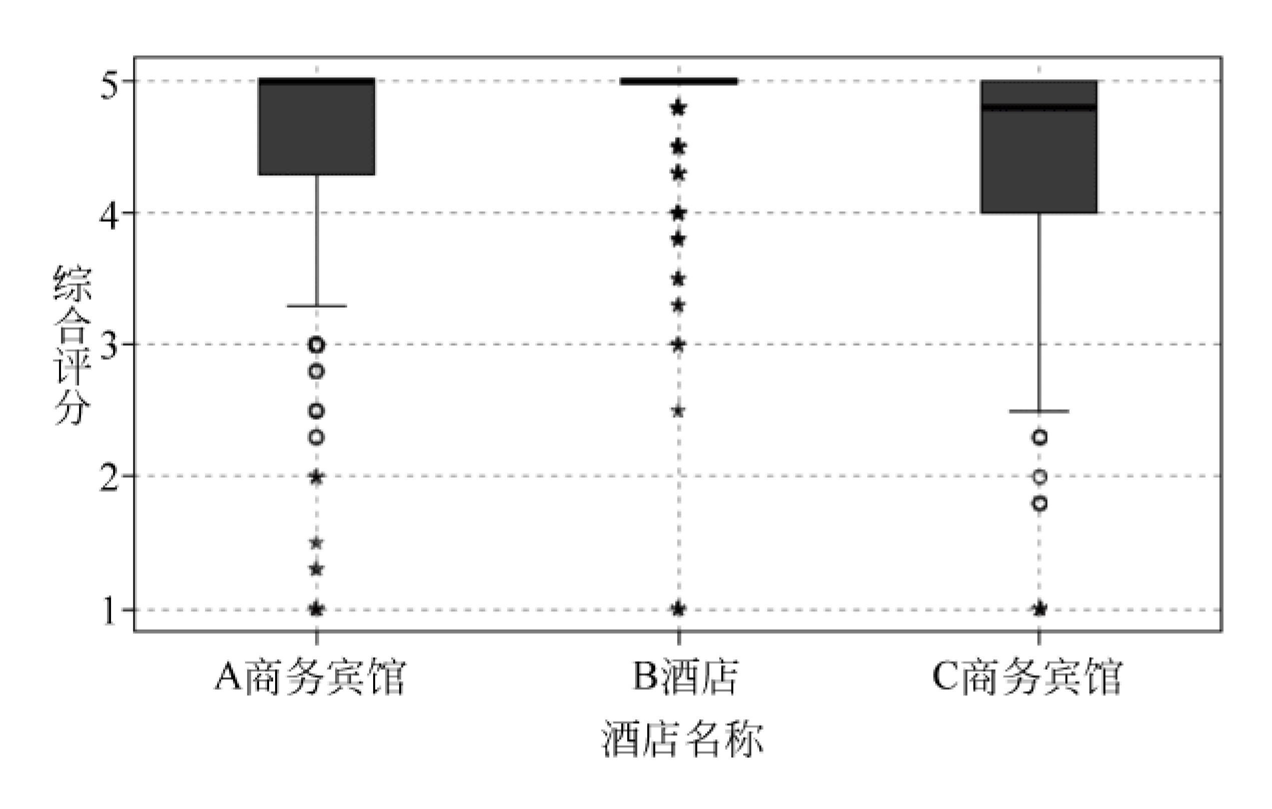


图 8.17 各酒店的综合评分箱图

从图 8.17 中可以看出酒店之间的评分差异较明显，为了分析酒店的评分与哪些因素相关，我们将评论的内容进行分词、去掉停用词、去掉“酒店”“房间”等被评价主体词汇，然后将得到的分词结果进行词频统计，获得关键词的排名和词频数，将其以标签云的形式显示出来，如图 8.18 所示。

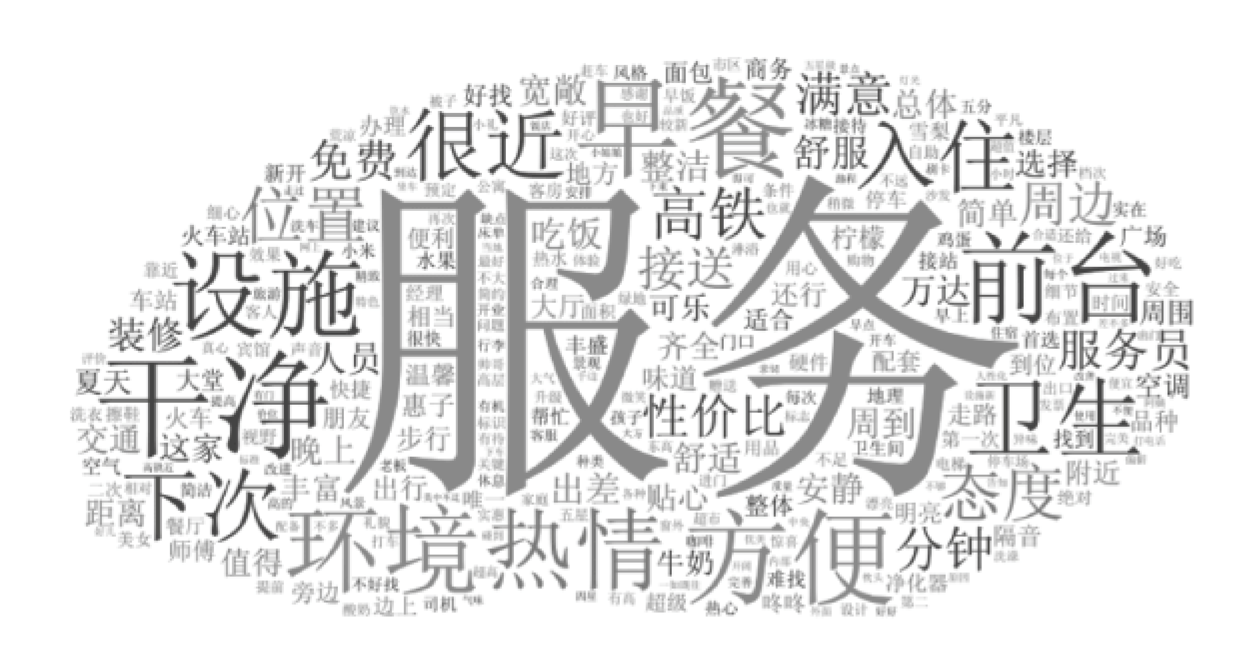


图 8.18 评论内容关键词标签云

从标签云中可以看出，客人最关心的因素是服务，其次是早餐、设施、干净、卫生、前台、环境、热情、方便、位置、态度、高铁等。

从图 8.19 中可以看出，出现早餐关键字的评论中，其评分的区分度较高，但是其与评分成反比，即评论出现了早餐字样，说明是负面评价较多。设施关键字并没有区分度，因为出现这一关键字评价中的各项评分几乎没有太多差别。干净和卫生为同一意义，其区分度基本一致，前台具有较高区分度，提到前台的评论中基本上是正面评价。

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.534	4.835	1.000 重要
位置评分	4.387	4.740	1.000 重要
设施评分	4.511	4.830	1.000 重要
服务评分	4.501	4.848	1.000 重要
卫生评分	4.688	4.899	1.000 重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.639	4.547	0.988 重要
位置评分	4.509	4.408	0.971 重要
设施评分	4.625	4.513	0.993 重要
服务评分	4.628	4.484	0.999 重要
卫生评分	4.756	4.728	0.595 不重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.622	4.642	0.375 不重要
位置评分	4.493	4.489	0.060 不重要
设施评分	4.601	4.646	0.659 不重要
服务评分	4.605	4.607	0.040 不重要
卫生评分	4.747	4.782	0.638 不重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.585	4.792	1.000 重要
位置评分	4.458	4.643	1.000 重要
设施评分	4.561	4.804	1.000 重要
服务评分	4.563	4.784	1.000 重要
卫生评分	4.717	4.899	1.000 重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.610	4.744	0.998 重要
位置评分	4.478	4.615	0.986 重要
设施评分	4.592	4.729	0.994 重要
服务评分	4.586	4.765	1.000 重要
卫生评分	4.743	4.826	0.960 重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.591	4.839	1.000 重要
位置评分	4.447	4.785	1.000 重要
设施评分	4.573	4.826	1.000 重要
服务评分	4.569	4.835	1.000 重要
卫生评分	4.729	4.896	1.000 重要

图 8.19　服务、早餐、设施、干净、卫生、前台关键字的双样本 t 检验结果

从图 8.20 中可以看出，环境、热情具有较高区分度，都为正面评价；而方便、位置、态度不具有区分价值，正面和负面评价均有；出现高铁关键字的评论中，对设施和卫生评分没有区分价值，对其他分项的评分虽然被标记为重要，但是其分值的区分度也不明显。

综上，具有较高的区分度的主要热词有：服务、早餐（负面）、干净、卫生、前台、环境、热情、高铁。

使用 CART 分类回归树模型对评论中的各分项评分以及用户级别、出行类别、房型等进行分类回归分析，结果如图 8.21 所示，从变量的重要性可以看出服务评分与综合评分具有较高的一致性，并且其重要性也最高，也与双样本 t 检验的结果相同。

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.596	4.801	1.000 ★重要
位置评分	4.463	4.680	1.000 ★重要
设施评分	4.578	4.787	1.000 ★重要
服务评分	4.573	4.812	1.000 ★重要
卫生评分	4.728	4.897	1.000 ★重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.585	4.967	1.000 ★重要
位置评分	4.444	4.917	1.000 ★重要
设施评分	4.566	4.963	1.000 ★重要
服务评分	4.560	4.996	1.000 ★重要
卫生评分	4.725	4.983	1.000 ★重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.629	4.594	0.653 不重要
位置评分	4.494	4.487	0.111 不重要
设施评分	4.613	4.571	0.673 不重要
服务评分	4.615	4.545	0.891 不重要
卫生评分	4.755	4.729	0.541 不重要

*单元格内容:平均值

字段	无关键字*	有关键字*	重要性
综合评分	4.623	4.643	0.290 不重要
位置评分	4.487	4.574	0.785 不重要
设施评分	4.607	4.601	0.072 不重要
服务评分	4.603	4.642	0.458 不重要
卫生评分	4.754	4.716	0.535 不重要

*单元格内容:平均值

字段	有关键字*	无关键字*	重要性
综合评分	4.673	4.621	0.611 不重要
位置评分	4.553	4.489	0.593 不重要
设施评分	4.675	4.603	0.703 不重要
服务评分	4.667	4.602	0.645 不重要
卫生评分	4.772	4.750	0.309 不重要

*单元格内容:平均值

字段	有关键字*	无关键字*	重要性
综合评分	4.568	4.639	0.971 ★重要
位置评分	4.402	4.518	0.995 ★重要
设施评分	4.560	4.620	0.889 不重要
服务评分	4.529	4.626	0.990 ★重要
卫生评分	4.732	4.757	0.589 不重要

图 8.20　环境、热情、方便、位置、态度、高铁关键字的双样本 t 检验结果

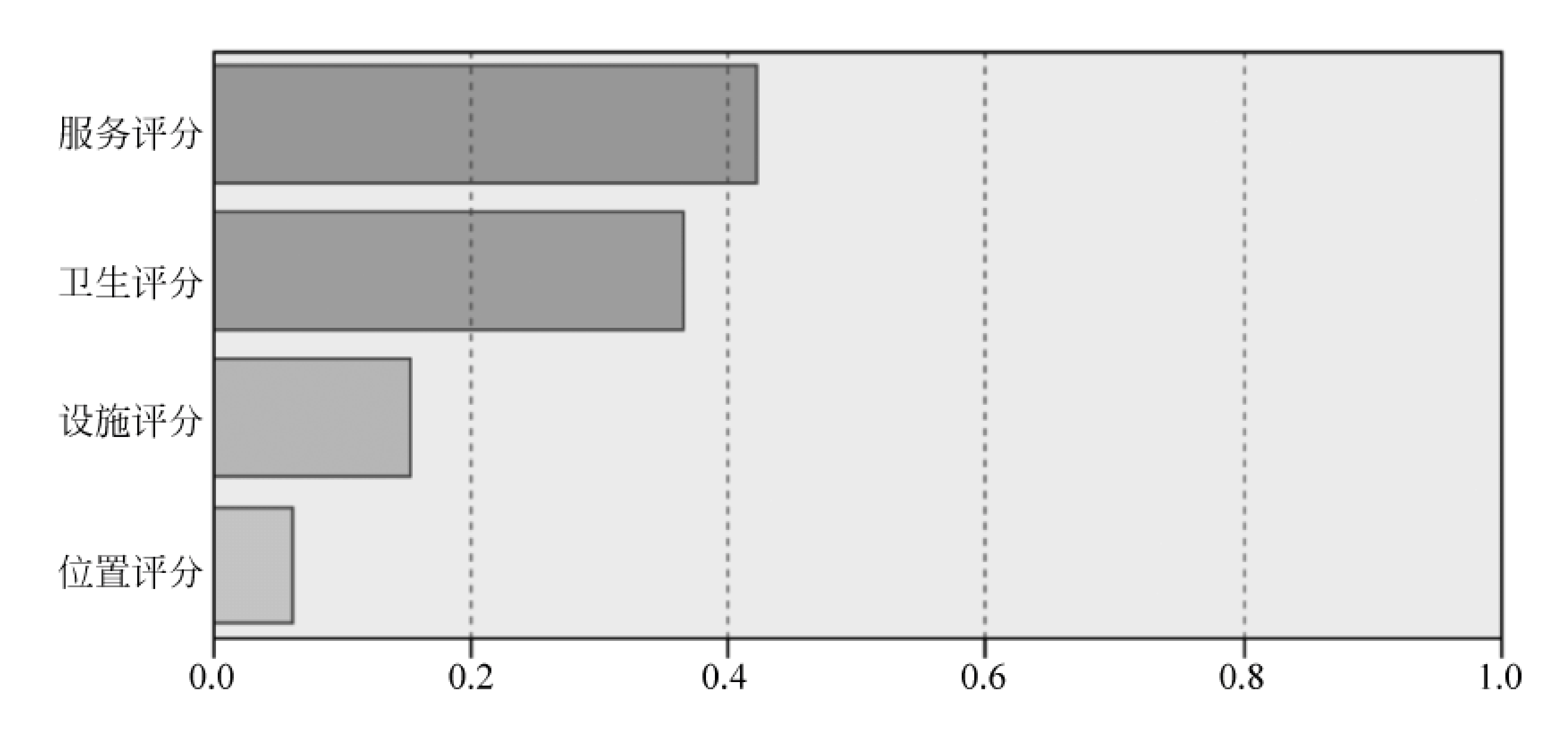


图 8.21　CART 分类回归树分析影响因素

对酒店评价高的一般情况下服务评分也较高，酒店卫生方面的重要性稍微次于服务，而酒店所在位置的影响因素较弱，即与酒店整体的评分关系并不强，可能是由于目前交通较为便利，对酒店地理位置的要求没有服务水平和卫生条件那么高，客人更重视入住之后的心理感受和卫生情况。

详细的分类结果如图 8.22 所示，服务评分以 4.5 为界限进行分类，低于 4.5 分的评论数为 457 条，占总数量的 27.431%，高于 4.5 分的评论数为 1209 条，占总数的 72.569%。其他各项评分的分类详情可查看图中标注。

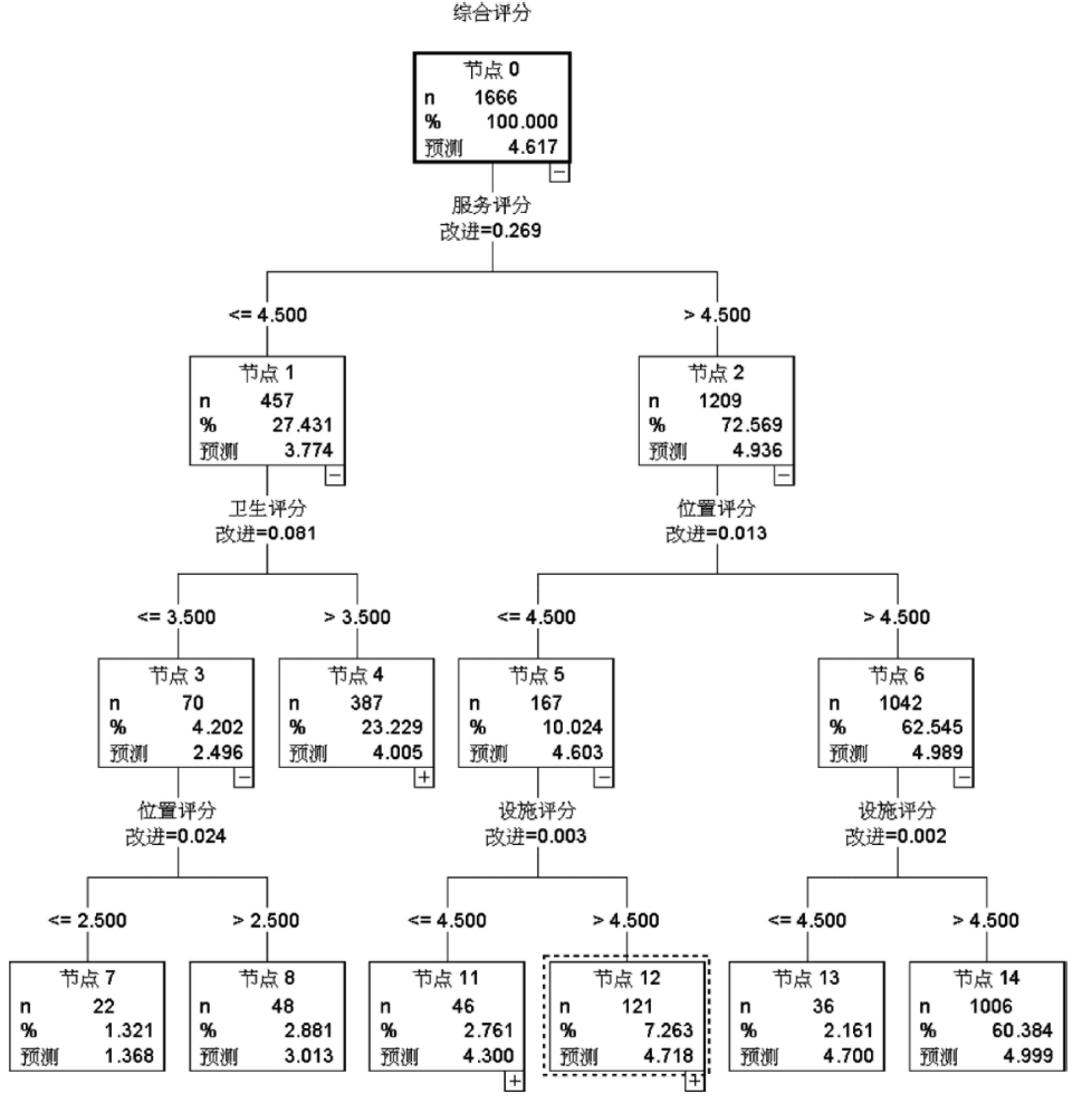


图 8.22　详细的分类结果

8.5.2　酒店评分与酒店业绩关系

在成本比较稳定的情况下，酒店的业绩收入主要由入住客人的数量决定，由于无法直接取得酒店中的实际经营数据，但是酒店的评论数量与酒店入住数量为正比关系，即通常情况下酒店的评论的数量较多时，入住客人数量也越多。所以，通过分析某一段时间内评论数量

和综合评分的走势情况来分析酒店评分对酒店业绩的影响。

首先，以周为统计单位计算酒店综合评分的平均分和评论条数总数，然后将评论的周平均分和周评论总数累加，形成累计值，这样可以从中看出总评分和总评论数随时间（周）变化的趋势情况。

图 8.23 是以时间散点图的形式显示 3 家酒店的评分和评论数走势，曲线的斜率说明了其数量增长速度，可以看到 A 和 B 两家酒店中，随着评分上升，评论数量呈更快速的增长，直观上说明评论分值可以对入住数量形成促进作用；而 C 商务宾馆在 2016 年 6 月之前走势基本与前两者相同，但是随着评分增长速度变缓，其评论数量的增长率越来越小，说明评分呈下降时，影响到酒店的业绩。

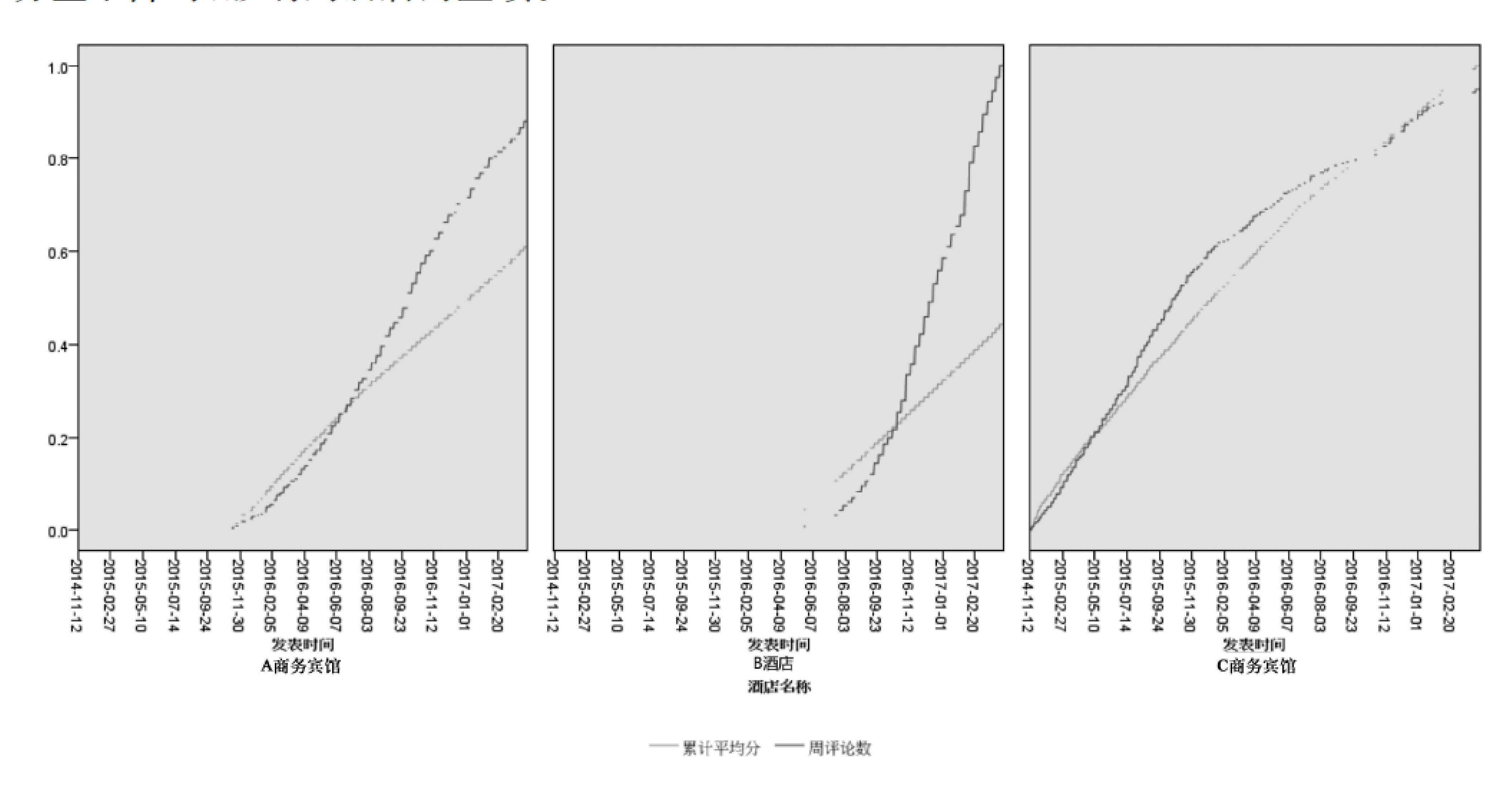


图 8.23 以周为单位的累计评分和累计评论数不同酒店对比

为量化累计综合评分和累计评论数量之间的关系，将两者应用于回归分析中，得到图 8.24 所示的结果，可以看出模型的增长曲线走势情况，3 家酒店的相关系数均达到 98% 以上。

图形	模型	酒店名称	相关
		B酒店	0.984
		C酒店	0.986
		A宾馆	0.991

图 8.24 回归分析累计评分与累计评论数关系

其中，B、C、A 酒店的 R 方值分别为 0.968、0.973、0.982，如图 8.25 所示，说明两者存在较强的相关关系。

不同酒店的回归分析相关系数结果如图 8.26 所示，分别对应 B、C、A 酒店。

综上，说明酒店评分可以直接影响酒店的经营业绩，如果综合评分显示较高，可以促进

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	0.984[a]	0.968	0.968	9.076677

a.Predictors:(Constant),周评论数

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	0.986[a]	0.973	0.973	23.525535

a.Predictors:(Constant),周评论数

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	0.991[a]	0.982	0.982	10.734467

a.Predictors:(Constant),周评论数

图 8.25　B、C、A 酒店以周为单位的累计评分和累计评论数回归结果

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	76.102	0.640		118.966	0.000
	周评论数	0.205	0.001	0.984	157.385	0.000

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	−3.922	1.692		−2.318	0.000
	周评论数	0.621	0.004	0.986	168.395	0.000

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
1	(Constant)	50.546	0.801		63.072	0.000
	周评论数	0.376	0.002	0.991	201.162	0.000

图 8.26　B、C、A 酒店回归分析相关系数

酒店客人的入住量，如果点评网站上的评分较差，可能会影响客人的印象，导致其不预订，最后影响酒店的业绩。

8.5.3　酒店评分分析

酒店基础分析包括评分趋势分析、房型分析、消费者决策分析等。酒店评分随时间的走势反映了酒店经营活动是否良好，客观反映酒店对问题的改进能力和适应变化能力，包括总评分及各分项评分的走势、按房型的评分走势分析等。

1. 酒店评分趋势

通过对酒店的评价分值按照时间维度进行分析，经过对评价数据进行按月度统计分析，将按照总平均分、位置评分、设施评分、服务评分、卫生评分进行统计，可以看到每个月的点评数量，间接可以看到数据的支持度。

所有房间类型随时间变化情况如图 8.27 所示，可以看到对酒店中所有房型的评分呈现波动性，评论时间是在入住之后由房客评价的，所以其并非是实际的入住时间，会有一定的时间差（几天到 1 个月之间不等），具有延后效应，所以在时间段上采用较模糊的时间段，并非严格对应自然月。

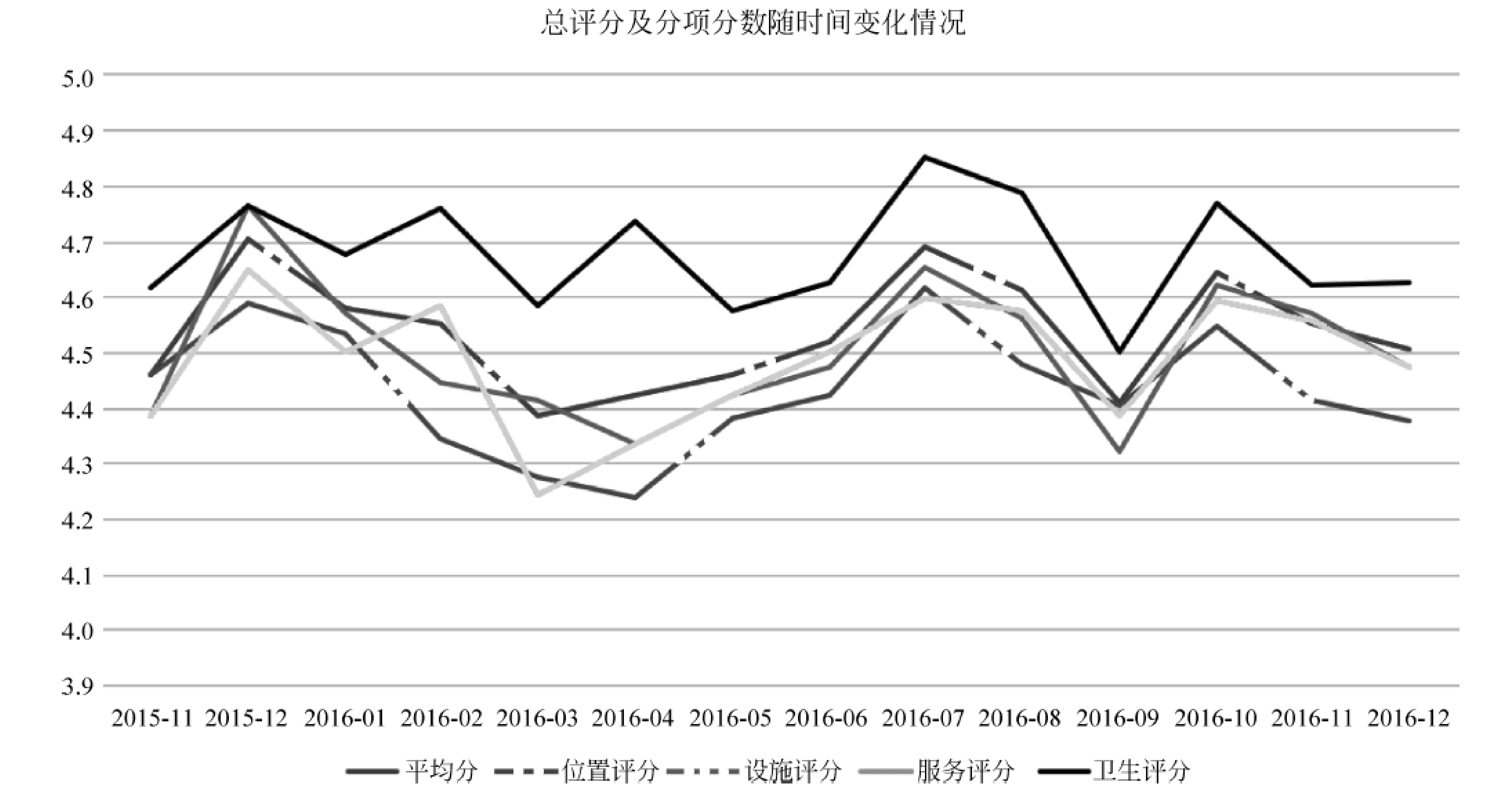


图 8.27 所有房间类型随时间变化情况

在这种情况下，虽然只有 659 条评论数据，但是依然可以看到 2016 年 3 月到 2016 年 4 月评分结果较差，而 2016 年 6 月到 2016 年 8 月之间的用户评价较高。从图 8.27 中可以看出，2016 年 3～4 月之后由于评分分数影响到了酒店经营，在其之后进行了服务质量改进，逐渐提高了客户满意度，如改善了早餐质量等。图 8.28 是分别将 2016 年 7 月和 2016 年 9 月的评论数据进行分词统计词频之后得到的词频标签云。由于“东站”“高铁站”等词作为酒店位置的描述，在问题描述中无实际意义，所以在生成标签云过程中将其去除。

7 月份的评价总体较高，原因是客人评分较低的原因中大部分是环境、位置不方便找、装修有味道等客观因素，而 9 月份评价较低的主要原因更多是人为的因素，如卫生间漏水、设施未及时维修等，并且此类问题在评论中的数量较多，一个月有 33 条评价，而 7 月份只有 17 条。从 9 月份的评论数据中可以看出管理上有明显的失误，随后经过调整，评分升高，但在后面 2 个月中评分又开始下降。从评分走势中可以看出，A 商务宾馆的管理和服务水平并不稳定，显示其有明显的管理失误和漏洞，原因可能是人员流失或服务标准过于随意。





图 8.28　2016 年 7 月(左)和 2016 年 9 月(右)的得分评价标签云

2. 房型分析

从房型上分析评分情况，将各个房型中对应不同评分项的平均分进行统计，结果见表 8.1。其中，温馨家庭房和观景双大床房评论数明显极少，只有 3 条评论，原因可能是这几种房型之间的区别较少，没有太高区分度，导致客人选择价格较低的舒适大床房，从表 8.1 中可以看出，舒适三人间的评价数为 15 条，样本数较少，其较高的 4.6 分值不作为参考，上述房型评论数量较少，将其剔除，以便能显示出其他各房型的评价差异程度。

表 8.1　A 商务宾馆房型、评分、评价数的对比

房　　型	综合评分	位置评分	服务评分	设施评分	卫生评分	评论数量
大床房	4.525 81	4.4409	4.4731	4.5161	4.6344	93
温馨家庭房	5.000 00	5.0000	5.0000	5.0000	5.0000	3
舒适三人间	4.600 00	4.4667	4.40	4.6000	4.8667	15
舒适双床房	4.594 08	4.5089	4.5503	4.5503	4.7278	169
舒适大床房	4.528 44	4.4094	4.5000	4.4875	4.6781	320
观景双大床房	3.933 33	3.6667	3.6667	4.0000	4.3333	3
阳光双床房	4.562 50	4.4375	4.4688	4.5625	4.7188	32
阳光商务大床房	4.539 13	4.5217	4.5217	4.4783	4.5652	23

从表 8.1 中的数据可以看出，舒适双床房评分好于舒适大床房，而舒适大床房的评论数量最多，达到 320 条，几乎是第二名舒适双床房的 2 倍，因为通常情况下，单人入住会优选大床房，所以可推断酒店的主要客户为单人入住居多，但是从表 8.1 中看到单人入住的评价分数却较差。另外，通过分析原始数据记录，在数据中综合评分低于 4.0 分的舒适三人间评论数为 15 条，舒适大床房的差评数量为 40 条，而对位置评分方面，舒适双床房为 14 条，舒适大床房为 47 条，宾馆中所有房型的地理位置其实是一样的，间接说明同样的服务水平、硬件设施情况下，舒适大床房的客人各方面的要求更高！

下面将列出舒适大床房和舒适双床房的出行类型和相对应的评分数值来观察出现此现象的主要原因。

将所有选择了舒适大床房的客人按照其出行类别进行分组，统计各出行类别下的各项评分分值，结果见表 8.2。

表 8.2 舒适大床房中各出行类别的评分对比

出行类别	综合评分	位置评分	服务评分	设施评分	卫生评分	评论数
代人预订	4.785 71	4.5714	4.8571	4.8571	4.8571	7
其他	3.800 00	3.6000	3.8000	4.0000	3.8000	5
商务出差	4.523 18	4.3906	4.5064	4.4893	4.6695	233
情侣出游	4.483 87	4.4839	4.4194	4.3548	4.6129	31
家庭亲子	4.700 00	4.6875	4.4375	4.6875	4.9375	16
朋友出游	4.745 83	4.6250	4.7917	4.7083	4.8333	24
独自旅行	3.650 00	3.2500	3.5000	3.2500	4.5000	4

表 8.2 中的“代人预订”“其他”“独自旅行”几类出行类别因数据太少不作分析，从表 8.2 中可以看出商务出差作为绝对的主流出行类别，情侣出游、家庭亲子和朋友出游远少于上述人群，分别只有 31 条、16 条、24 条，将这几类出行类别的各项评分进行可视化，结果如图 8.29 所示。

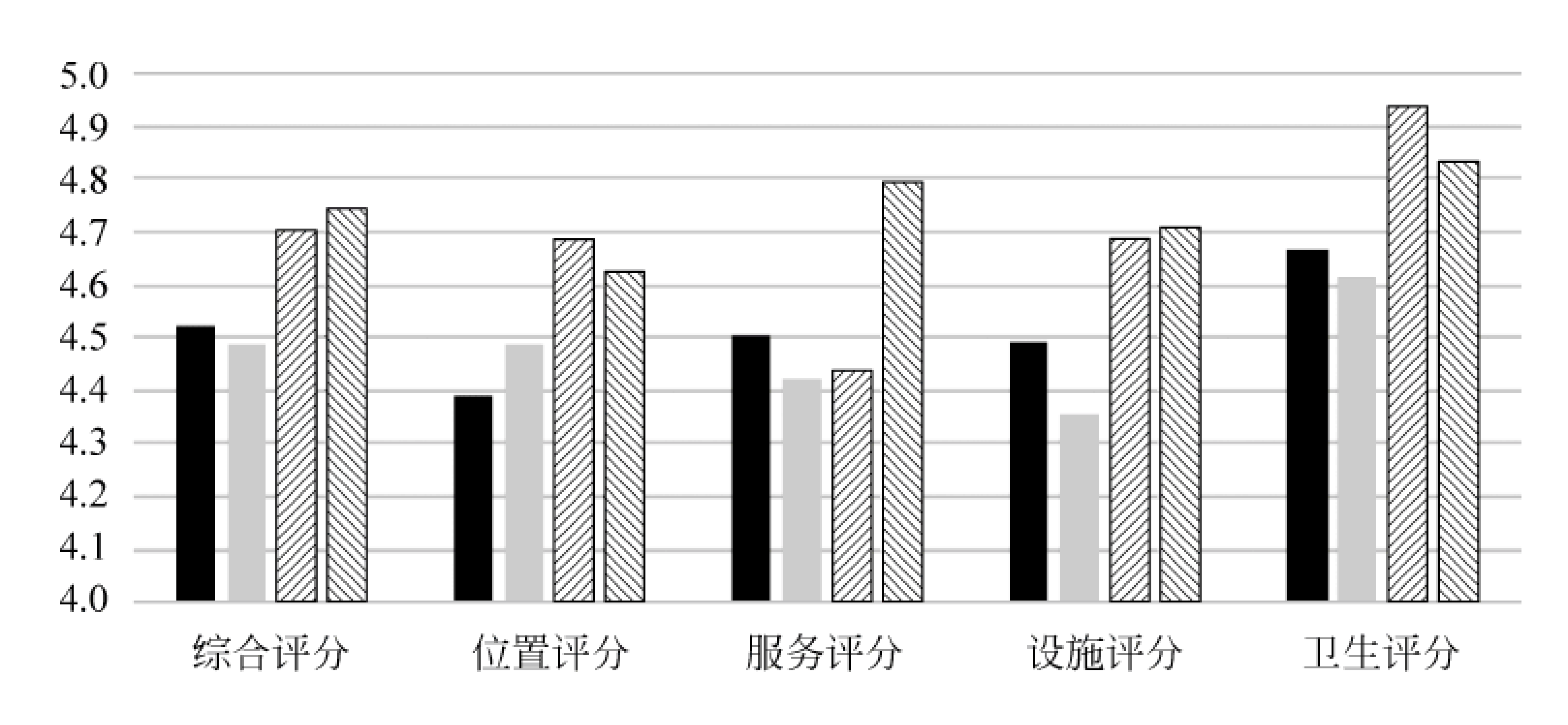


图 8.29 舒适大床房中不同出行类别评分对比

虽然情侣出游人群的评论条数较少，但从中可以发现这类人群选择了舒适大床房后对各方面的要求高于商务出差人群，除位置评分外，包括综合评分在内的其他各项评分都为最低评分，这一人群对服务和设施的要求较高(实际评分较低)，建议宾馆前台人员对此类选择舒适大床房的人群进行特别照顾，相应提高服务标准，否则会拉低宾馆的整体评分。

将所有选择了舒适双床房的客人按照其出行类别进行分组，统计各出行类别下的各项评分分值，结果见表 8.3。

表 8.3 舒适双床房中各出行类别的评分对比

出行类别	综合评分	位置评分	服务评分	设施评分	卫生评分	评论数
代人预订	5.000 00	5.0000	5.0000	5.0000	5.0000	1
其他	3.750 00	4.0000	3.5000	3.5000	4.0000	2
商务出差	4.548 76	4.4215	4.5124	4.5124	4.7025	121
情侣出游	4.875 00	5.0000	4.7500	4.7500	5.0000	4
家庭亲子	4.690 00	4.6667	4.6667	4.6667	4.7333	30
朋友出游	4.757 14	4.8571	4.5714	4.5714	5.0000	7
独自旅行	5.000 00	5.0000	5.0000	5.0000	5.0000	4

表 8.3 中主流人群为商务出差和家庭亲子两类人群，情侣出游人群一般不会选择双床房，所以在表 8.3 中只有 4 条评论内容，在分析中予以剔除，图 8.30 是可视化结果。

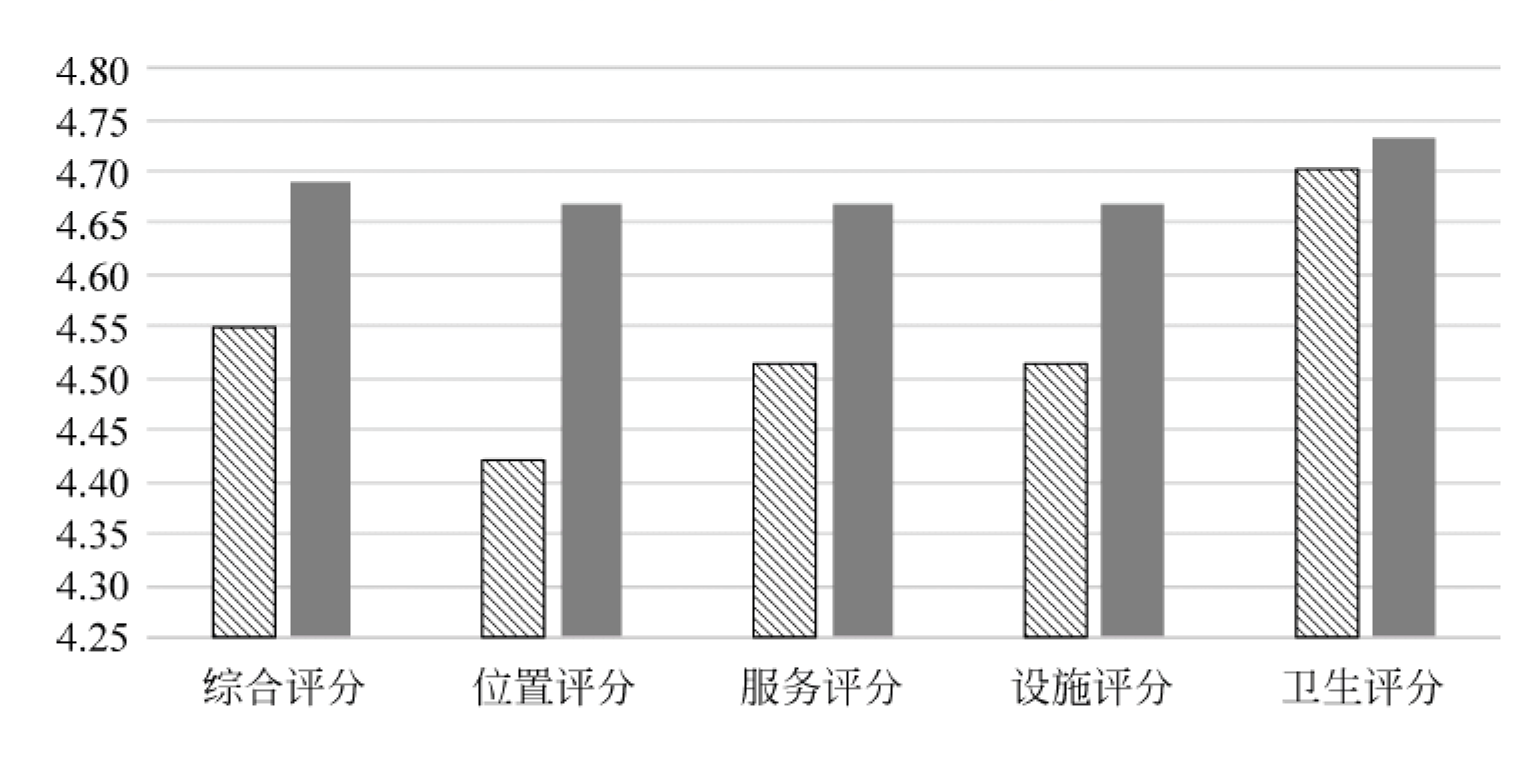


图 8.30　舒适双床房中不同出行类别评分对比

从图 8.30 中可以看出，在舒适大床房的入住人群中，商务出差人士的包括综合评分在内的所有评分都较低，说明这类人群对宾馆的各方面要求较高，目前 A 商务宾馆并没有达到其要求，只是符合家庭出游类人群的要求层次，后文将分析家庭出游人群其实是酒店入住较少的人群，所以说明 A 商务宾馆的经营服务水平尚在商务型酒店的初级阶段。

以箱线图的形式显示不同房型的评分情况，如图 8.31 所示，左侧是箱线图，由于其分值的分布可能是较少的评论数产生的，使结果有较大误差，容易产生不稳定的干扰，所以需要配合二维点图一起分析，右侧是不同房型的二维点图，表示不同房型的情况下，不同评分值的数据分布情况。从图 8.31 中可以看出，温馨家庭房虽然分数极高，但是数据条数却极少，最终其结果不具有参考价值。

综上，面对情侣出游和商务出差两类客人时，需要特别注意细节，提高服务水平，必要时提供客户关怀等服务。另外需要注意的是，房型过于集中在极少数房型中，导致其区分度不高，不利于充分利用所有房型，容易使某些房型空置率增高，而另外的房型紧张，也间接说明了房间设置不合理，没有对不同人群匹配不同的房型，这种情况下也难以实现差异化服务。

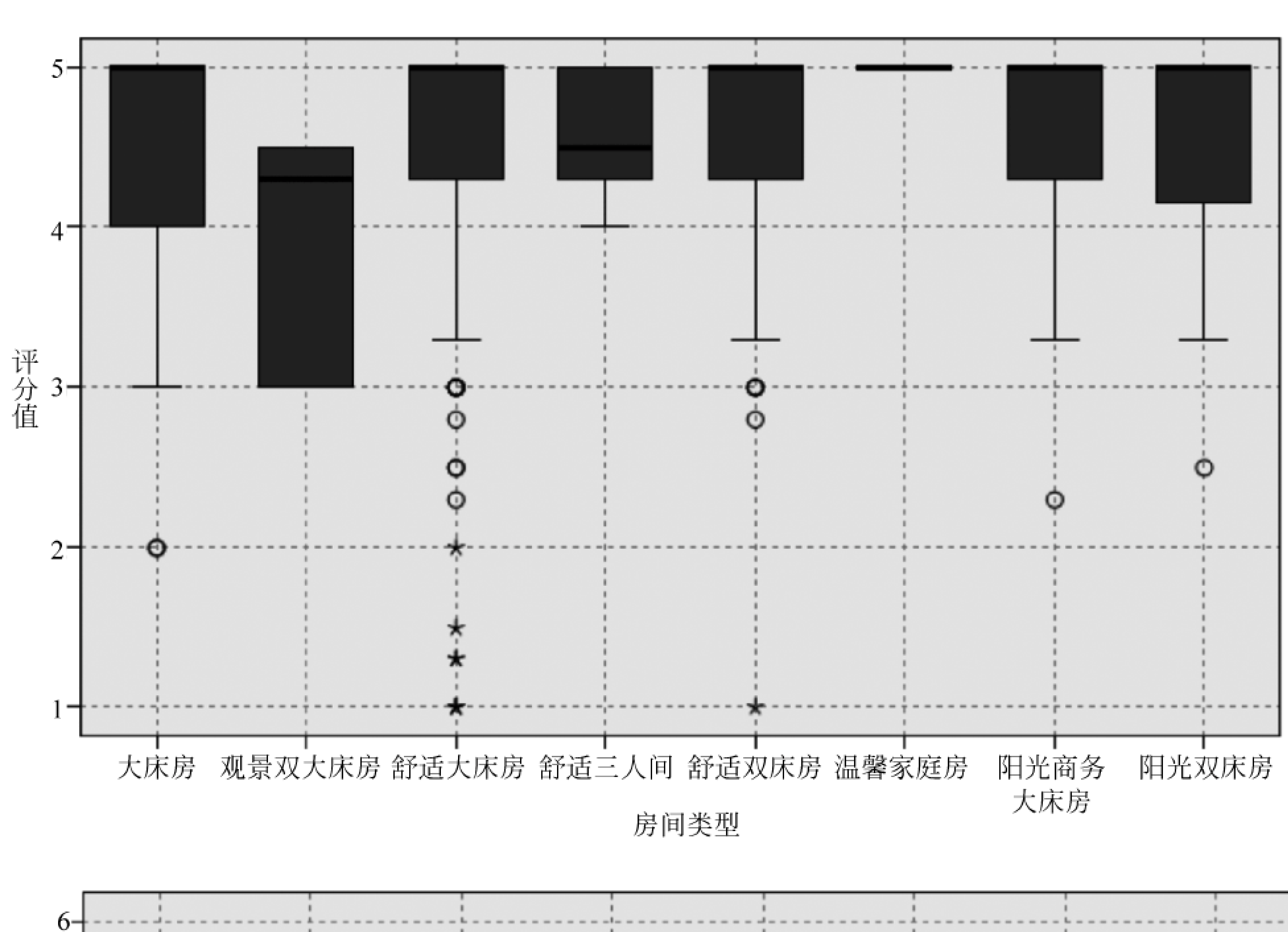
3. 消费者决策分析

对消费者的分析有助于了解其需求并制定相应营销、服务等策略，包括消费者画像、消费特征分析、客户忠诚度分析等。

通过对评价内容中顾客选择的出行类别分析，获取基本的用户画像数据，将对不同类别的入住客人进行评分数值进行分析，以便分析 A 商务宾馆的主要适用人群。

1）按出行类别分析

按照出行类别对入住客人进行统计，表 8.4 列出了不同的出行类别对应的评价数量和总平均分情况。可以看出商务出差人群是 A 商务宾馆的主要客户群体，与其他经济型酒店基本一致，其次是家庭亲子群体，图 8.32 是对数据的可视化展示，可以直观看出其数量对比情况。



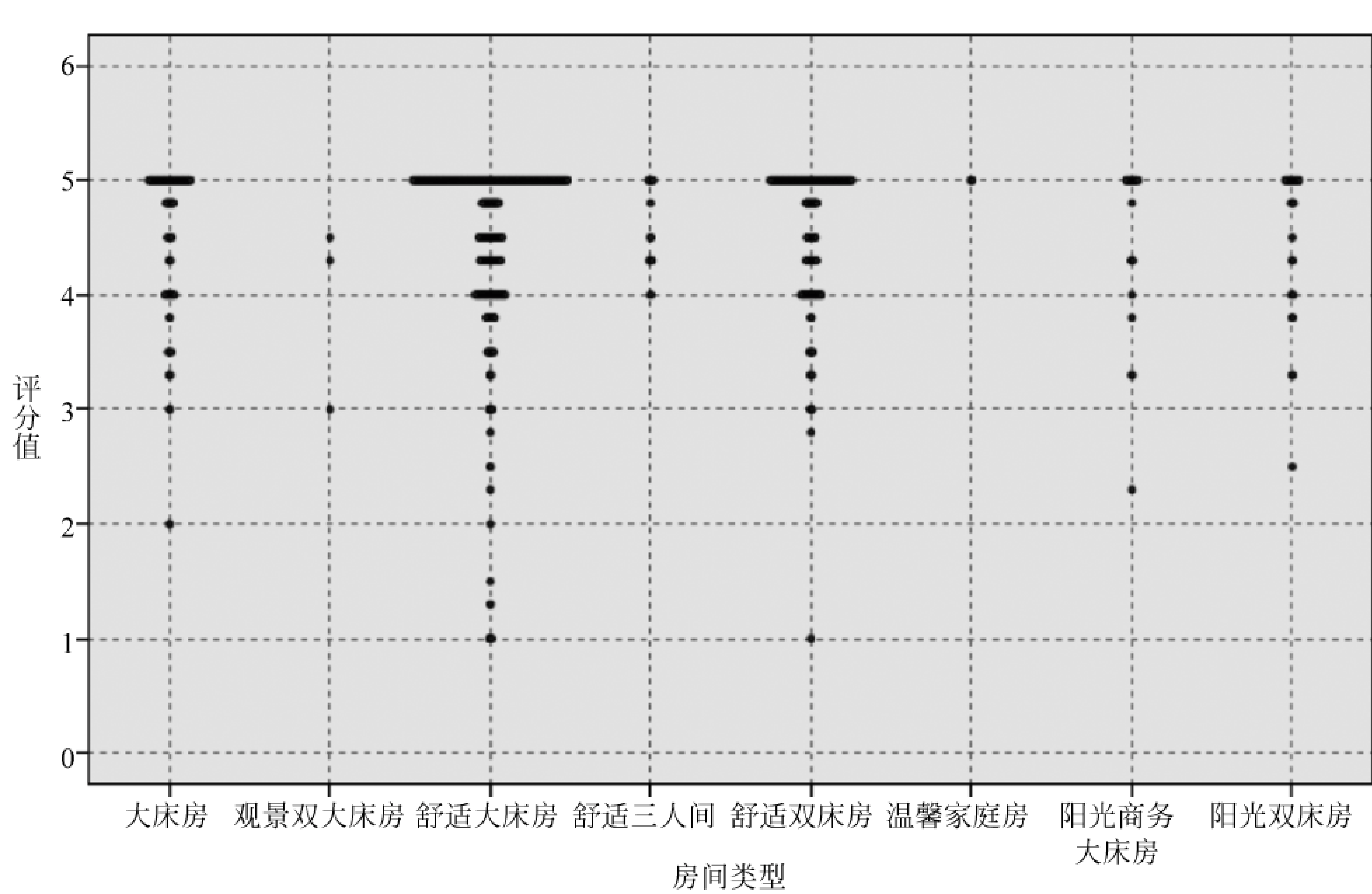
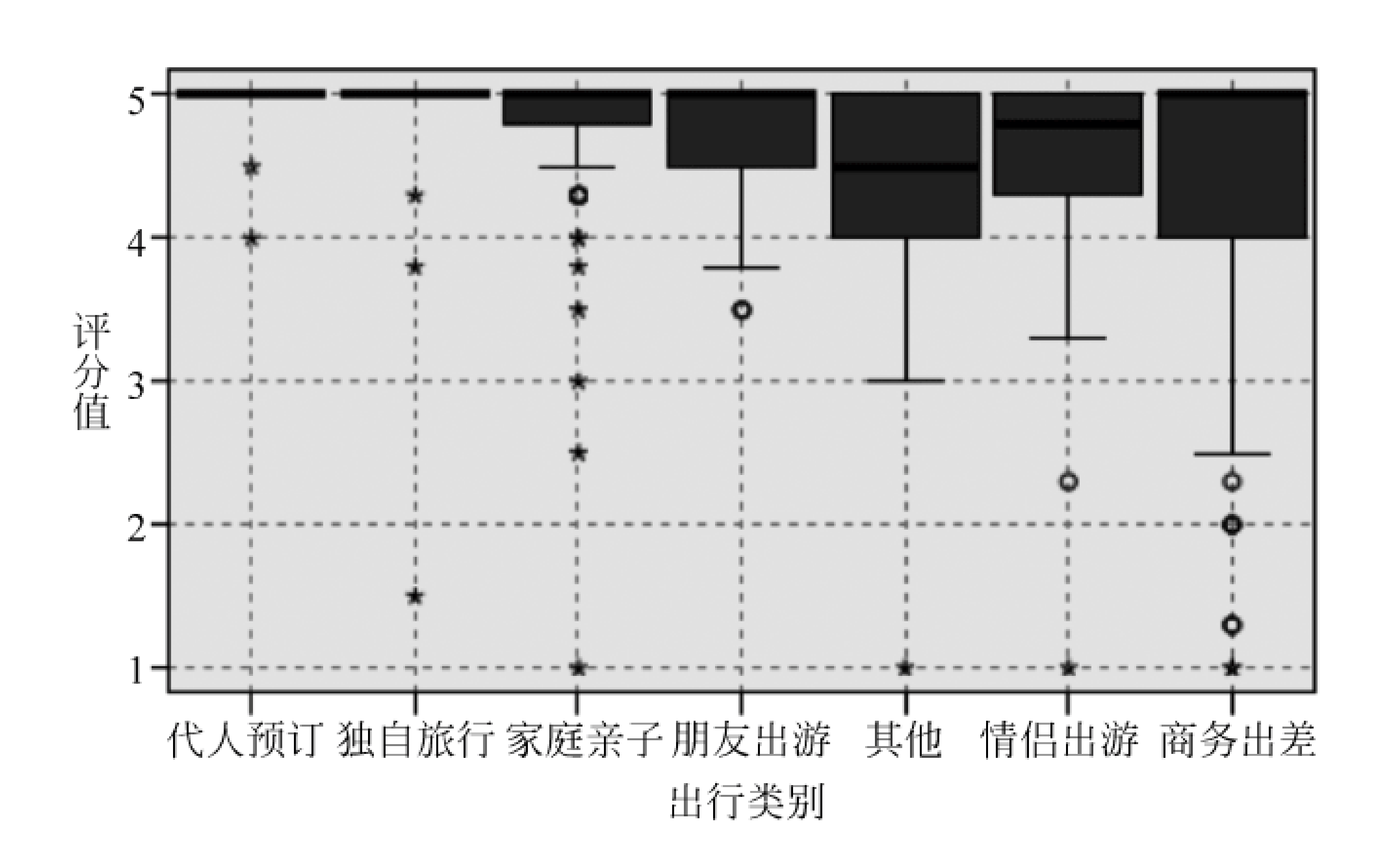


图 8.31 按房间类型的评分情况，上方是箱图，下方是二维点图

表 8.4　出行类别对应数量和评分

出行类别	总平均分	数量
代人预订	4.833 33	9
其他	4.100 00	10
商务出差	4.519 54	476
家庭亲子	4.711 76	68
情侣出游	4.462 50	40
朋友出游	4.700 00	39
独自旅行	4.662 50	16



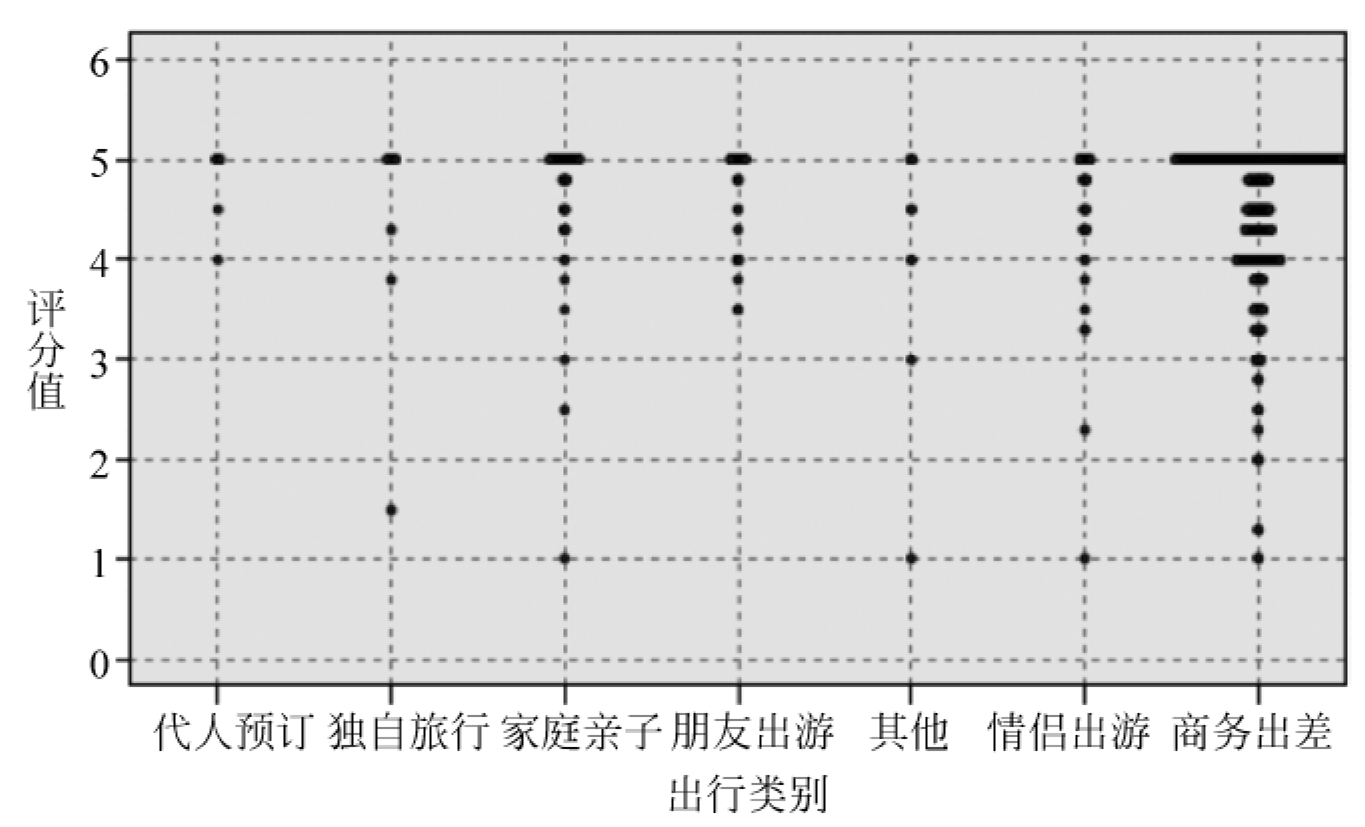


图 8.32　不同出行类别对应的入住数量和评分情况

由于代人预订这一类别的用户没有真实体验，只是听被代订人员的评价，一般情况下会对评价有一定的美化效果，并且其评价数量只有 9 条，不具有代表性，暂不作为分析依据，“其他”类别的评论数较少，也不作分析。在其余各类出行类别中，商务出差和情侣出游的评分最低，家庭亲子类别的评分最高。从箱图中可以看出商务出差人群的评分范围最广，低分区间明显低于其余各类别，二维点图（下侧）显示了各评分人群的数量，从 5.0 分到 4.0 分之间并非依次递减的顺序，而是打分为 4.0 分的仅次于 5.0 分，说明在商务出差人群中对 A 商务宾馆的整体评价较低。

原因可能是在同样情况下，商务出差和情侣人士对细节和服务水平要求更高，其满意度较低也符合常理，但商务人群正是商务酒店的主要服务人群，如果不能在此类用户中产生较好的满意度，则会严重影响酒店的盈利。另外，消费者中家庭亲子类总体评价较高，说明对这类人群而言，宾馆的定位符合其预期，后续可对这部分客人重点进行推广。总之，说明 A 商务宾馆对主要的客户人群方面、服务方面表现不尽如人意，核心竞争力较差。

2）不同类别客人的评价分析

按照客人在评价网站上全部评论条数对客人进行分类，点评数超过 30 条的称为“点评专家”，5～29 条是“点评达人”，低于 5 条的称为“点评新星”，点评数量越多说明入住酒店次数多，具有较多的入住经历，评价相对更加客观，同时对酒店的各方面要求也会较高。图 8.33 是 3 类人群对酒店的整体评分情况。

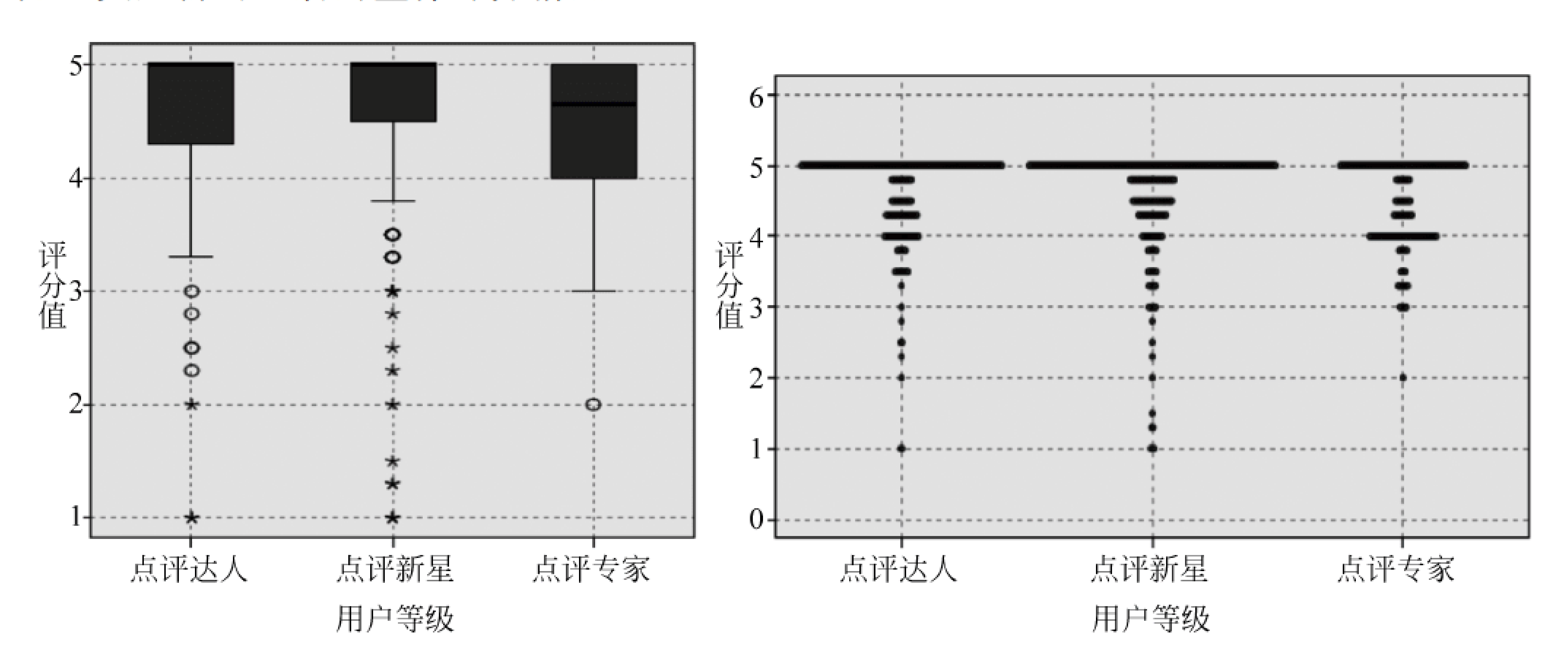


图 8.33 A 商务宾馆不同等级客人评分情况

从图 8.33 中可以看出点评专家和点评达人对酒店的评价总体不高，分值较为分散，并且 4.0 分的数量明显高于 4.0～5.0 分的数量，而点评新星对 A 商务宾馆的评价较高，分布也较集中，4.0～5.0 的分数分布情况依次递减。说明 A 商务宾馆在初次入住人群中具有较好的口碑，但是在后续住过其他同类酒店之后，经过比较，可能不会再选择入住，意味着其在红海市场的竞争中具有一定的弱势。

对不同级别的客人选择出行类别进行交叉分析，如图 8.34 左所示，点评专家绝大多数为商务出差，也与实际相符。点评新星除商务出差外，较多为家庭亲子出行，如首次外出旅游。点评达人也与之类似，只是在家庭亲子的数量上较少。图 8.34 右是不同级别客人对房型的选择情况，从中可以看出点评专家类商务出差大部分情况是入住舒适大床房，间接说明这类客人一般为单独出差较多，可有针对性推荐相关产品或服务。其他级别的客人与点评专家基本一致，只是入住大床房的数量略多一些，说明其他两类人群对价格略微敏感。

另外也可以看出在不同等级的客人之间，房型选择上趋于相同，间接说明酒店房型设置上存在问题，因为不同的客人对入住需求不同，可有针对性地提供更多差异化服务或房型进行创新。

3）客人关注点

将客人提交的评论进行分词，并对词频进行计算，去掉类似“酒店”等无意义词汇，按照问题的重复数量进行排序，对存在的问题进行比较，如图 8.35 所示。

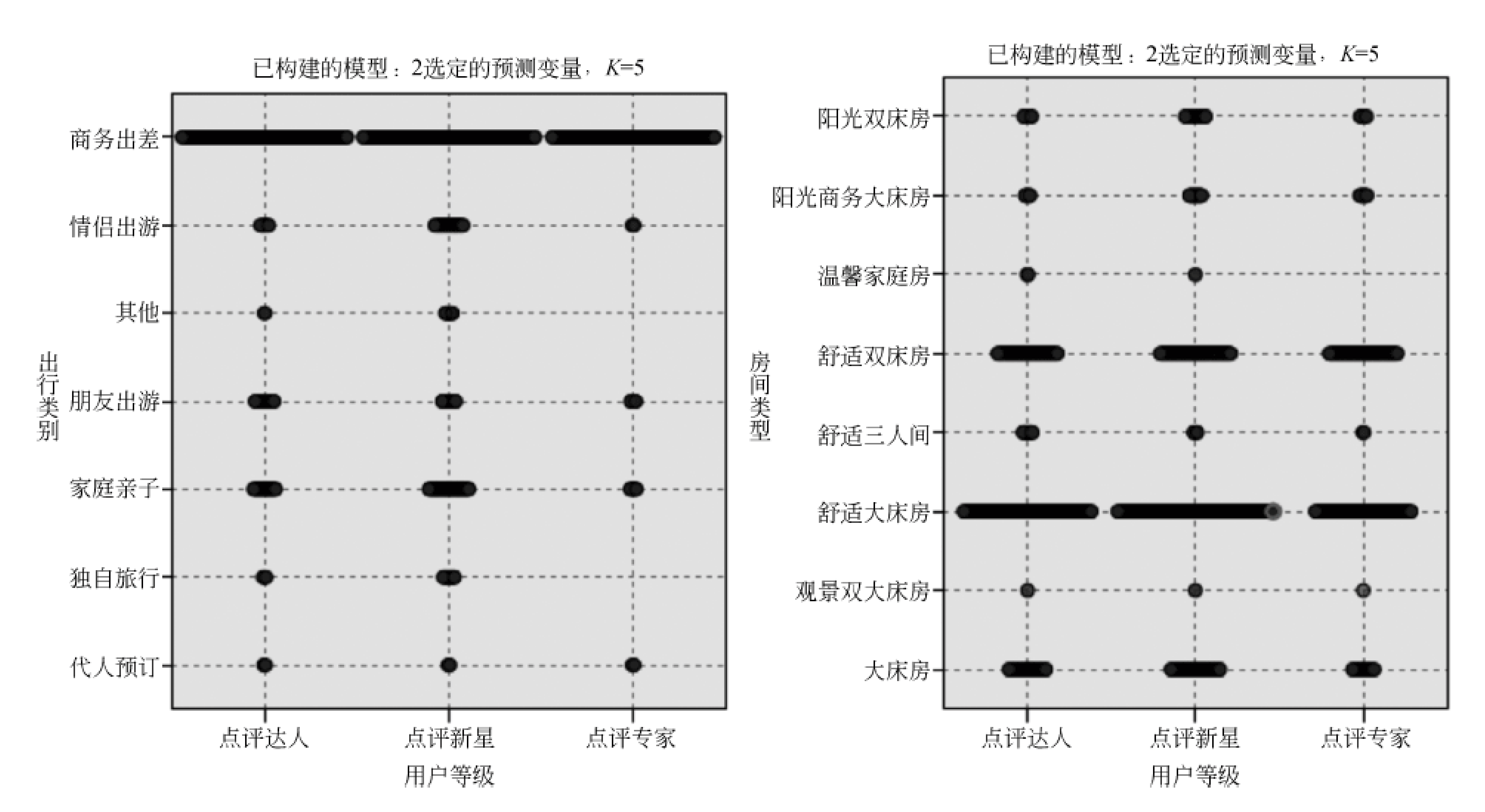


图 8.34　A 商务宾馆不同等级客人出行类别和房间选择情况

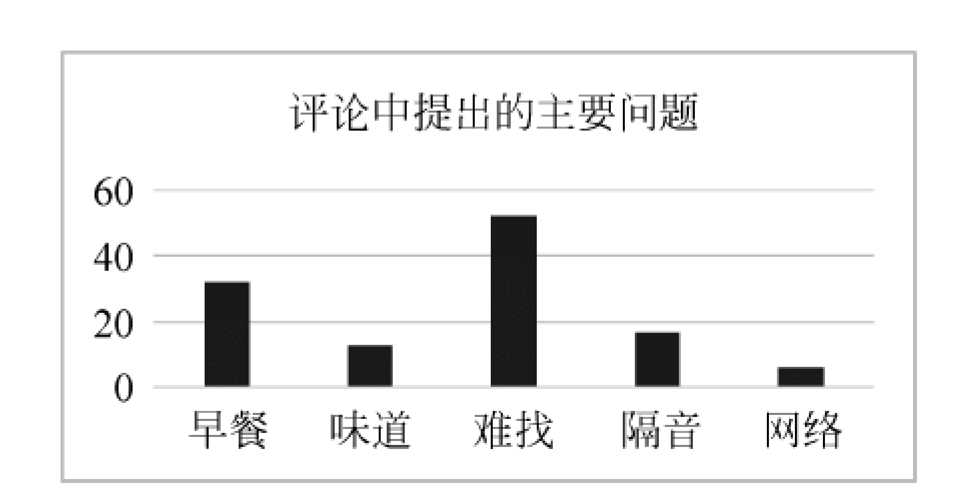


图 8.35　A 商务宾馆评论中提取的主要问题数量对比

其中，早餐除了难吃之外，还有客人评价品种太少。在所有评价中，涉及餐饮的达到 122 条，占总评论数的 19%，说明客人对饭食的关注较高。味道是指房间中有装修的味道，难找是指酒店位置在导航软件中不易查找，总相关条数达到 52 条，占总数的 8%，评价中指出隔音较差 17 条，占总评价数的 3%，网络问题提出数量为 6 次。其他问题还包括周边配套、环境等客观因素，由于改进能力有限，不作分析。

4. 客户忠诚度分析

用户的复购率反映了其对品牌的忠诚度，在本次分析中的计算方式是重复入住（以点评为准）人数与总人数之比，复购次数多说明其为酒店的忠诚会员。表 8.5 是复购超过 2 次（大于 2 次）的客户列表，人数是 21 人，总人数为 541 人，超级复购人群占比为 3.88%，比率较低，说明非常愿意入住本酒店的人数较少，酒店的品牌效应不明显，依然没有形成自主的客户资源。

通过查询样本数据，发现复购 2 次以上的人数（包括 2 次）的总人数为 69 人，占总人数的比例为：69 人/541 人＝12.75%，对于商务型酒店来说，这样的复购比率说明回头客较少，品牌的忠诚度不够。

B 酒店的复购率是 137 人/886 人＝15.46%，而 C 商务宾馆的复购率是 114 人/845 人＝13.49%，另外两家酒店的复购率均超过 A 商务宾馆，说明 A 商务宾馆在客户关怀或服务方面有待改善。

表 8.5 复购超过 2 次的用户

昵 称	购买数量	昵 称	购买数量
118543 ****	3	past1231	12
118639 ****	3	todaye	3
288018 ****	3	yc ****	3
300251 ****	4	ZC66 ****	8
320041 ****	3	品味人生	4
320516 ****	4	夜光小曲	4
Aren861013	6	张天扬 2006	4
E5360 ****	4	战神甲骨文	3
M32817 ****	3	舒妹	6
M46774 ****	3	鞠鸿印	4
M6925 ****	3		

8.5.4 客户情感分析

通过对 A 商务宾馆客户评论内容进行情感分析,得到顾客对酒店各方面的态度和情感倾向,获得客户更加关注哪些属性,从而建立用户体验模型,形成用户对酒店服务各方面的关注权重;同时,对于酒店比较欠缺的方面,提出针对性的改进建议。

希望通过对用户评论数据的分析,挖掘出用户对该酒店的整体情感倾向。由于语言数据的特殊性,主要是将一篇句子中的关键词提取出来,从而将一个评论的关键词也提取出来,然后根据关键词所占的权重,应用空间向量的模型,将每个特征关键词转化为数字向量,通过计算其距离,得到聚类,从而得到情感的分类,用来表示客户的情感倾向。

首先将原始评论数据集中的评论数据单独转存为 txt 文本格式,然后在进行分词之前,先对数据的基本情况进行审查,发现评论数据具有以下特点:

(1) 大多数评论数据情感倾向比较明显,涉及情感的关键词比较集中。

(2) 评论数据不规范,存在一些网络词、表情符号等。

(3) 评论数据之间存在重复的现象,特别是单条评论中往往存在一些重复的词语;可能是用户评论的时候直接复制、粘贴其他人的评论内容。

(4) 标点符号比较多。

为使得评论数据达到符合情感分析的标准,对其进行三级清洗:一级清洗(去除标点符号)、二级清洗(去除重复内容)、三级清洗(去除停用词、网络用词等)。

在信息检索中,为节省存储空间和提高搜索效率,在处理自然语言数据(或文本)之前或之后会自动过滤掉某些字或词,这些字或词被称为停用词(Stop Words)。对这些停用词需要进行必要的处理,通过整合现有的停用词库,包括“百度停用词库”“哈尔滨工业大学停用词词库”“四川大学机器学习智能实验室停用词库”,形成了一个更加完善的停用词词库,其中包含 1980 个停用词。

一句话中出现的重复词汇也会影响到一个评论中关键词在整体中出现的词频,从而影响整体的分析结果,所以要对其进行压缩。

基于情感词典的方法,需要用到标注好的情感词典。本案例中,直接使用知网发布的

“情感分析用词语集”，使用武汉大学开发的 ROST CM6 作为中文分析的工具，分词之前如图 8.36 所示。

```
设施还不错，挺干净，就是晚上走廊脚步声太多！挺棒的！离高铁又近，就是早餐太晚，赶高铁都吃不到早餐了！
房间非常干净整洁，宽敞明亮，服务也很好，离高铁站很近，非常方便，赠送的早餐也很丰富，推荐入住！
"不错吹吹噢，不错up的，不错的啊。。。。。。
离火车徐州东站相当的近啊。走路十分钟啊。相当的号啊。棒"
干净的酒店，出门就是高铁。晚上零点楼层异响持续半小时，打了两次电话才解决，实在影响睡眠。除了这点都还不错。
房间不错很安静，美中不足就是周边吃饭的地方太少了，如果临时换车住一下的话，还是不错的选择，性价比超高值得推荐
离徐州东站非常近，走几分钟就到。室内打扫挺干净的，房间够大，比较宽敞，卫生间也是。洗澡很舒服。早餐还可以。
非常好的商务酒店，绿地开发的房子第一印象就不错，设施新，干净，环境好，就在高铁站旁边。绝对好评
离高铁站近，方便
吃饭要到高铁吃快餐，周边没有饭店。酒店离高铁也就三四百米远。
这是我第二次住这酒店了，高铁出口往右直行穿过高架桥下小路不到200米就到了。房间面积大，干净整洁，周边不喧闹。
非常方便，离高铁站很近，房间宽敞整洁。标间还有1.5米的两个床，也有1.2米床的房间。来回住了两晚。下次还会再去住
酒店干净，设施齐全，房间大，更重要的是服务员热情。早餐我们过点了，还热心地帮我们热早餐。
真的不错?
挺干净，房间也不错
只有早餐，其他都完美
房间不错，高铁站边上，客服人员非常热情，微笑服务，感觉很温馨。就是夏天的被子太厚了。
周边配套设施还有待完善!酒店不太好找。
离火车站步行15分钟时间，房间中规中矩，空调直吹身上有点受不了
房间很干净，装修也不错
离高铁站很近，坐车方便。卫生环境以及设备也都挺好，服务态度也不错。洗发水洗发液都是一小瓶，有点少，还缺少个吹
三张床放着也不觉得挤，室内简洁干净，非常不错。
环境很好，下次继续入住
环境优雅，视野开阔，房间又大又干净，下次一定还要入住
酒店硬件设施很好，房间大，干净卫生，无线比较快，赶火车没有吃早餐，离徐州东站很近，步行五分钟，赶车方便。
进去一定要注意台阶！其余都OK高铁很近，价格也可以接受
环境不错。不错不错的。
房间比较大，设施较新，离高铁站很近，就是一进屋子很大味道
```

图 8.36　原始数据集

经过分词处理之后的结果数据集如图 8.37 所示，部分词语被强制分开，如“异响”等，为此使用自定义词表，使这类词汇不作分词。

```
设施 还 不错 ， 挺 干净 ， 就是 晚上 走廊 脚步声 太 多 ！ 挺 棒 的 ！ 离 高铁又 近 ， 就是 早餐 太 晚 ， 赶
房间 非常 干净 整洁 ， 宽敞 明亮 ， 服务 也 很 好 ， 离 高铁站 很 近 ， 非常 方便 ， 赠送 的 早餐 也 很 丰富
" 不错 吹吹 噢 ， 不错 up 的 ， 不错 的 啊 。 。 。 。 。 。
离 火车 徐州 东站 相当 的 近 啊 。 走路 十 分钟 啊 。 相当 的 号 啊 。 棒 "
干净 的 酒店 ， 出门 就是 高铁 。 晚上 零点 楼层 异 响 持续 半 小时 ， 打 了 两 次 电话 才 解决 ， 实在 影响
房间 不错 很 安静 ， 美中不足 就是 周边 吃饭 的 地方 太 少 了 ， 如果 临时 换车 住 一下 的话 ， 还是 不错 的
离 徐州 东站 非常 近 ， 走 几 分钟 就 到 。 室内 打扫 挺 干净 的 ， 房间 够 大 ， 比较 宽敞 ， 卫生间 也 是
非常 好 的 商务 酒店 ， 绿地 开发 的 房子 第一 印象 就 不错 ， 设施 新 ， 干净 ， 环境 好 ， 就 在 高铁站
离 高铁站 近 ， 方便
吃饭 要 到 高铁吃 快餐 ， 周边 没有 饭店 。 酒店 离 高铁 也就 三四 百米 远 。
这 是 我 第 二次 住 这 酒店 了 ， 高铁 出口 往 右 直行 穿过 高架桥 下小 路不 到 200 米 就 到 了 。 房间 面积
非常 方便 ， 离 高铁站 很 近 ， 房间 宽敞 整洁 。 标 间 还有 1. 5 米 的 两 个 床 ， 也 有 1. 2 米 床 的 房间
酒店 干净 ， 设施 齐全 ， 房间 大 ， 更 重要 的 是 服务员 热情 。 早餐 我们 过 点 了 ， 还 热心 地 帮 我们
真的 不错 ?
挺 干净 ， 房间 也 不错
只有 早餐 ， 其他 都 完美
房间 不错 ， 高铁站 边上 ， 客 服 人员 非常 热情 ， 微笑 服务 ， 感觉 很 温馨 。 就是 夏天 的 被子 太 厚 了
周边 配套 设施 还有 待 完善 ！酒店 不 太 好找 。
离 火车站 步行 15 分钟时 间 ， 房间 中 规 中 矩 ， 空调 直 吹 身上 有点 受不了
房间 很 干净 ， 装修 也 不错
离 高铁站 很 近 ， 坐车 方便 。 卫生 环境 以及 设备 也 都 挺 好 ， 服务 态度 也 不错 。 洗发水 洗发 液 都 是
三 张 床 放 着 也 不觉 得 挤 ， 室内 简洁 干净 ， 非常 不错 。
环境 很 好 ， 下次 继续 入住
环境 优雅 ， 视野 开阔 ， 房间 又 大 又 干净 ， 下次 一定 还要 入住
酒店 硬件 设施 很 好 ， 房间 大 ， 干净 卫生 ， 无线 比较 快 ， 赶 火车 没有 吃 早餐 ， 离 徐州 东站 很 近 ，
进去 一定 要 注意 台阶 ！ 其余 都 OK 高铁很 近 ， 价格 也 可以 接受
环境 不错 。 不错 不错 的 。
房间 比较 大 ， 设施 较 新 ， 离 高铁站 很 近 ， 就是 一 进 屋子 很 大 味道
```

图 8.37　经过分词处理之后的结果数据集

为确定用户评论数据集中包含的情感类型，采用聚类对其进行分析，分别尝试了分为 3 类和 4 类的聚类，最终发现分为 3 类的情绪更加合理，于是将用户评论数据中表现出的情绪分为积极情绪、中性情绪、消极情绪三类。

基于上面的分析，使用武汉大学开发的 ROST CM6 对用户的评论数据进行情感分析，

发现用户的情感倾向见表 8.6。

表 8.6 情感分析占比表格

情绪倾向	情 绪 得 分	评论条数	百分比
积极情绪	一般(0～10)	314	46.87%
	中度(10～20)	173	25.82%
	高度(20 以上)	56	8.36%
中性情绪	—	102	15.22%
消极情绪	一般(－10～0)	21	3.13%
	中度(－20～10)	3	0.45%
	高度(－20 以下)	1	0.15%

从上面的分析可知,对该酒店整体情况持积极态度的占比大约是 81.04%,持中立态度的用户占比为 15.22%,持消极态度的用户占比 3.73%。其中,持积极态度和中立态度的用户总体占比为 96.27%,和网站提供的数据 97%非常接近,如图 8.38 所示。



图 8.38 网站对 A 商务酒店的分析数据

除了得到用户评论中的情感倾向之外,还希望了解每类用户在表达情感的同时所关注的因素,无论是酒店房间、地理位置、卫生,还是服务、餐饮等方面,希望通过挖掘持有正面态度的用户关注的因素找到酒店当前做得比较好的地方,以便后续在这些方面进一步加强,形成自己的招牌特征;同时希望通过挖掘持有中立态度的用户关注的因素得到酒店为了在未来吸引更多的用户所需要改进加强的地方;最后,希望对剩下少部分持有负面情感的用户关注的因素进行仔细分析,希望发现他们对酒店的不满意之处。

首先对分词处理后的用户评论数据进行词频统计,得到如图 8.39 所示的情况。

将"东站""高铁站"等词去掉后,借助标签云对该部分的统计结果进行可视化呈现,如图 8.40 所示。

通过对用户评论的整体分析发现,该酒店的客户比较关注的方面有:方便、早餐、环境、干净、设施以及服务等。结合前面对该酒店的基本分析,也可以发现该酒店位于高铁站附近,有非常大的地理优势,方便了很多出行的客户。同时,其环境和服务业较好,性价比较高,是旅游出行非常好的选择。

1. 正面情感用户分析

对上一步情感分析得到的正面情感用户的评论数据集进行进一步分析,首先通过分词、词频统计得到如图 8.41 所示的词频分布情况,其中,"高铁站""东站"等停用词已经去掉。

高铁站	119	周边	24	人员	15	选择	10	唯一	7	被子	5	大楼	5
方便	113	吃饭	24	出差	15	安全	10	标志	7	楼下	5	沙发	5
干净	102	高铁	23	总体	15	适合	10	高层	7	招牌	5	第一次	5
环境	84	交通	23	值得	15	二次	10	百米	6	免费	5	首选	5
设施	76	整洁	22	服务员	13	配套	9	高铁很	6	面积	5	品种	5
早餐	70	味道	22	晚上	13	满意	9	整体	6	周到	5	开阔	5
服务	59	距离	21	旁边	12	周围	9	好几	6	建议	5	入口	4
徐州	52	舒服	20	出行	12	绿地	9	相当	6	家庭	5	大厅	4
入住	49	宽敞	20	空调	12	卫生间	9	也就	6	刷卡	5	硬件	4
卫生	48	还行	19	热情	12	效果	8	条件	6	声音	5	开心	4
东站	45	好找	18	舒适	11	快捷	8	牌子	6	接待	5	早饭	4
性价比	40	隔音	18	难找	11	商务	8	小时	6	高的	5	集团	4
下次	35	火车站	17	视野	11	时间	7	超级	6	还要	5	位于	4
分钟	35	态度	17	齐全	11	这家	7	五分	6	完美	5	稍微	4
位置	34	附近	17	火车	11	便利	7	朋友	6	当地	5	办理	4
前台	33	地方	16	早上	10	地理	7	预定	6	种类	5	明显	4
步行	26	车站	16	宾馆	10	新开	7	手边	6	高铁近	5	赶早	4
安静	26	找到	15	走路	10	实惠	7	打电话	5	超市	5	坐车	4
装修	24	边上	15	简单	10	淋浴	7	楼层	5	便宜	5	接受	4

图 8.39　分词后统计结果



图 8.40　分词后可视化呈现

通过对上述词频统计结果进行可视化，得到如图 8.42 所示的标签云，可以看出持有正面情感的用户比较关注的因素主要是干净、方便、设施、环境、早餐、性价比、卫生情况、位置、交通、周边等。

这部分因素与整体上客户关心的因素基本上一致，是吸引消费者的因素，建议该酒店在后续的发展中要不断加强这些因素的竞争力，形成自己的特色，为更多的消费者提供更方便、更优质的服务。

干净	94	装修	18	出行	11	唯一	6	被子	5	品种	5	简单	4
环境	77	宽敞	18	晚上	11	配套	6	免费	5	商务	5	简洁	4
设施	71	舒服	15	适合	10	牌子	6	超级	5	门口	4	市区	4
方便	61	隔音	15	空调	10	五分	6	周到	5	大楼	4	十七	4
早餐	59	值得	15	周围	9	好几	6	地理	5	大厅	4	高铁的	4
服务	54	态度	14	绿地	9	也就	6	家庭	5	阿姨	4	高楼	4
卫生	47	车站	14	二次	8	实惠	6	淋浴	5	超市	4	高的	4
性价比	37	总体	14	火车	8	标志	6	当地	5	位于	4	感谢	4
前台	32	好找	14	早上	8	高铁很	6	还要	5	打电话	4	热水	4
入住	32	人员	13	难找	8	这家	6	开阔	5	稍微	4	银杏树	4
下次	30	出差	13	安全	8	高层	6	相当	5	饭店	4	再次	4
位置	30	味道	13	卫生间	8	条件	6	整体	5	办理	4	集团	4
步行	23	附近	13	快捷	8	手边	6	面积	5	招牌	4	马路	4
周边	23	找到	13	走路	8	百米	6	接待	5	小时	4	发票	4
安静	21	边上	12	地方	8	便利	6	舒适	5	赶早	4	洗澡	4
整洁	21	服务员	12	齐全	8	效果	6	选择	5	心情	4	有待	4
吃饭	21	视野	11	宾馆	8	时间	6	朋友	5	醒目	4	高铁近	4
交通	20	旁边	11	还行	7	楼下	5	沙发	5	影响	4	早饭	4
距离	18	热情	11	新开	7	第一次	5	楼层	5	建议	4	种类	4

图 8.41 正面情感分词后统计结果



图 8.42 正面情感分词后可视化结果

2. 中立情感用户分析

通过对情感分析阶段得到的中立情感用户评论数据进一步分词、统计词频分布，得到如图 8.43 所示的词频分布情况。

进一步使用标签云的方式对统计分析结果进行可视化，结果如图 8.44 所示。

从可视化标签云的呈现结果可以看出：对于持中立态度的用户而言，他们关注的情况主要在于酒店的方便性、酒店附近的交通情况、周围环境、早餐质量、房间装修情况、吃饭等。由此可见，该酒店后续有待进一步提升的服务包括早餐的口味、房间的装修情况，特别是有

方便	6	出差	2	吹风机	1	便捷	1	外地人	1	标示	1
还行	6	淋浴	2	将近	1	违章	1	不出	1	标志	1
环境	5	边上	2	吵醒	1	女朋友	1	入住	1	隔音	1
装修	4	车站	2	干净	1	不满足	1	系统	1	相当	1
早餐	4	距离	2	劣质	1	分钟时	1	找到	1	即可	1
附近	3	简单	2	效果	1	凑合	1	被套	1	集成	1
位置	3	甲醛	2	漏水	1	家具	1	广告	1	枫叶	1
分钟	3	火车	2	早上	1	棒棒	1	差不多	1	积水	1
交通	3	地理	2	总体	1	关键	1	小时	1	捡漏	1
预定	3	便宜	2	总的	1	下回	1	房地	1	唯一	1
设施	3	地方	2	建议	1	下来	1	坐车	1	进来	1
下次	3	商务	2	卫生	1	网上	1	各项	1	厕所	1
气味	3	种类	1	地处	1	尚可	1	舒适	1		
步行	3	超标	1	龙头	1	缘故	1	头发	1		
吃饭	3	整体	1	出站口	1	普通	1	注意	1		
味道	3	空调	1	无用	1	不通	1	卫生间	1		
性价比	2	郊区	1	无标	1	火车站	1	吊顶	1		
姑姑	2	硬件	1	开车	1	二次	1	服务	1		
走路	2	过道	1	身上	1	不明	1	梧桐	1		

图 8.43　中立情感分词后统计结果

图 8.44　中立情感分词后可视化结果

的用户在评论中指出该酒店的早餐有点不合口味。另外，房间由于装修存在甲醛味道，这都是该酒店在后续的经营过程中需要改进的地方。

3. 负面情感用户分析

通过对持有负面情感的用户的评论数据的分析，得到如图 8.45 所示的词频分布情况，从中可以看出早餐问题、服务态度、宾馆位置难找等问题较突出。

早餐	4	分钟	1	特意	1	进攻性	1	查验	1	入口	1	某某	1
入住	4	完成	1	明明	1	效果	1	测试	1	大堂	1	连接	1
态度	3	不像	1	明显	1	位置	1	环境	1	大不	1	单据	1
难找	3	老板	1	宾馆	1	翻翻	1	楼道	1	卡住	1	主管	1
服务	2	后悔	1	设施	1	操作	1	骗人	1	品目	1	一同	1
配套	2	降温	1	招牌	1	周边	1	素质	1	打电话	1	齐全	1
实在	2	结账	1	银行	1	还行	1	夜里	1	有问题	1	商务	1
刷卡	2	这家	1	银杏	1	还给	1	浴巾	1	果真	1	跟头	1
员工	2	遇到	1	刚刚	1	这次	1	停机	1	小时	1	紧张	1
地方	2	隔音	1	提示	1	早上	1	停水	1	朋友	1	开车	1
听到	2	隔壁	1	所在	1	发给	1	下车	1	尴尬	1		
装修	2	靠边	1	电视	1	发发	1	说不	1	规模	1		
下次	2	性价比	1	服务员	1	洗脸	1	师傅	1	失误	1		
人员	2	超级	1	不了了	1	记得	1	二十	1	客房	1		
完了	2	上次	1	蹦蹦车	1	水平	1	二次	1	物业	1		
房时	2	沟通	1	方法	1	太太	1	浓重	1	物品	1		
晚上	2	空调	1	包子	1	前台	1	和稀	1	车库	1		
滚滚	2	差劲	1	别来	1	无奈	1	满意	1	火车	1		
千万	1	旅馆	1	干净	1	说白	1	忙活	1	七点	1		

图 8.45　负面情感分词后统计结果

通过标签云的方式可视化，得到图 8.46 所示的结果。持有负面情感的用户关注更多的是酒店的房间、早餐情况、服务等，并且有用户还指出酒店难找，可能是该酒店没有设置明显的指示牌，导致部分用户难以找到酒店；另外，还有部分用户说该酒店的装修情况影响到其居住感受，结合中立用户的评论情况，该酒店可能最近新装修，所以有部分用户反映房间中依旧存在甲醛味道，所以这一点酒店要注意改进。

图 8.46　负面情感分词后可视化结果

在负面情绪中指出的装修甲醛问题，酒店应该给予充分的重视。倘若真的存在新装修空气不好的问题，势必会对客源造成极大的冲击。因此，酒店应该及时检测空气质量，保证

客源不会受到影响。

为了更加直观地展示各关键词标签之间的关联关系，以话题的方式显示用户关注问题，对情感分词后的结果进行网络语义分析，使用 RostCM 内集成的 NetDraw 软件构建语义网络，其构建过程是将分词后的结果文件生成高频词表，并将词表进行过滤，将无意义词汇滤除，得到过滤后的结果表；接着进行特征提取，获得行特征词表；最后通过对高频词进行共现分析得到共现矩阵，将其输入到 NetDraw 软件中绘制语义网络，结果如图 8.47 所示。

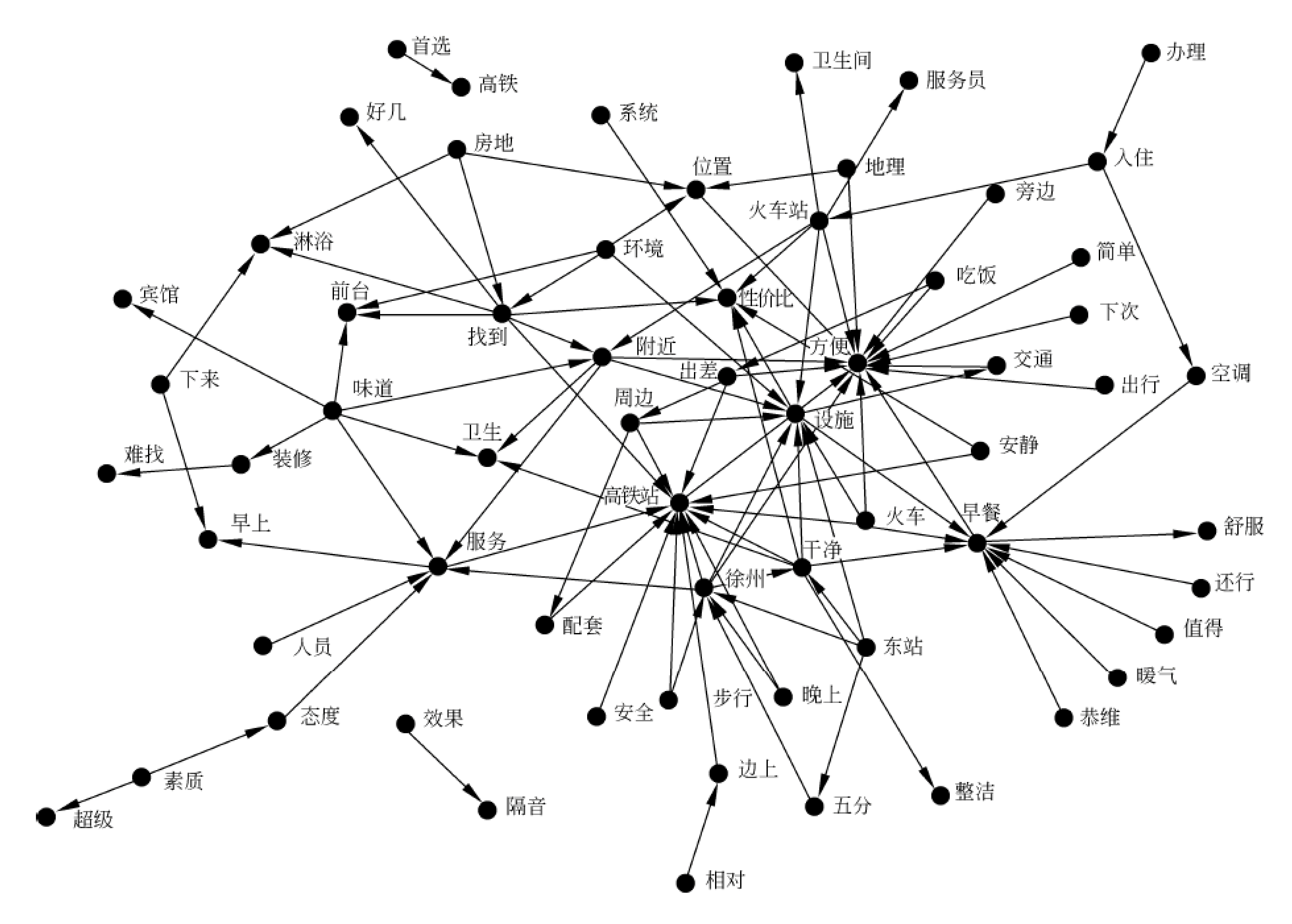


图 8.47　用户关注焦点语义网络

从有向图网络可以看到其中各分词结果之间的关联关系。通过分析关键节点的连接数量和方向，从中分析出酒店评价结果相关的主要影响因素及其之间的关系。

从图 8.47 中可以看出位置(高铁站)、设施、方便、性价比、早餐、服务等连接数较多，并且均为被指向的方向，说明这几个因素是影响酒店评价的重要因素。除此之外的附近、找到、味道等也可以侧面看出它们对酒店的评价产生了部分负面影响。外围节点中显示的是对核心影响因素的修饰，如“交通”是“方便”的，“早餐”是(不敢)“恭维”的，“装修”是有“味道”的等，建议后续将此节点中具有较多连入关系的影响因素点进行相应的干预，使其优点得到加强，缺点得到改进，逐渐增加较高评价因素的节点，减少负面评价的节点数。

8.5.5　竞争分析

虽然 A 商务宾馆附近存在数十所同样价位的商务宾馆，由于差异化不明显，竞争者都采用相同的经营管理模式，随着宾馆数量越来越多，店均客源逐渐减少，对 A 商务宾馆的经营产生较大影响，A 商务宾馆要想在竞争中立于不败之地，不仅需要了解自己存在的问题，更要对比分析其竞争优势和客源吸引能力，使酒店在经营过程中知己知彼，最终在竞争中逐

渐胜出。

选取 A 商务宾馆周边的商务宾馆和星级酒店作为竞争对手进行竞争力分析，它们分别是：D 商务酒店、B 酒店、C 商务宾馆，其中 B 酒店的定位是星级酒店，定位略有不同，其余两家为经济型商务酒店，为直接竞争对手，从点评网站上抓取了上述各家酒店的评论数据进行对比分析。

1. 客户评分对比

网站上的酒店评分是客人进行实际消费之后打的分数，相对较客观，对于未住过此酒店的客人来说，具有较强的决策影响作用，是酒店实力和竞争力的一种体现。通过对各商务酒店的综合评分和各分项评分进行比较，可得出其基本的竞争力情况。

2. 综合评分比较

获取了网站爬虫数据后，将 4 家对比酒店的综合点评平均得分进行了对比，结果如图 8.48 所示。

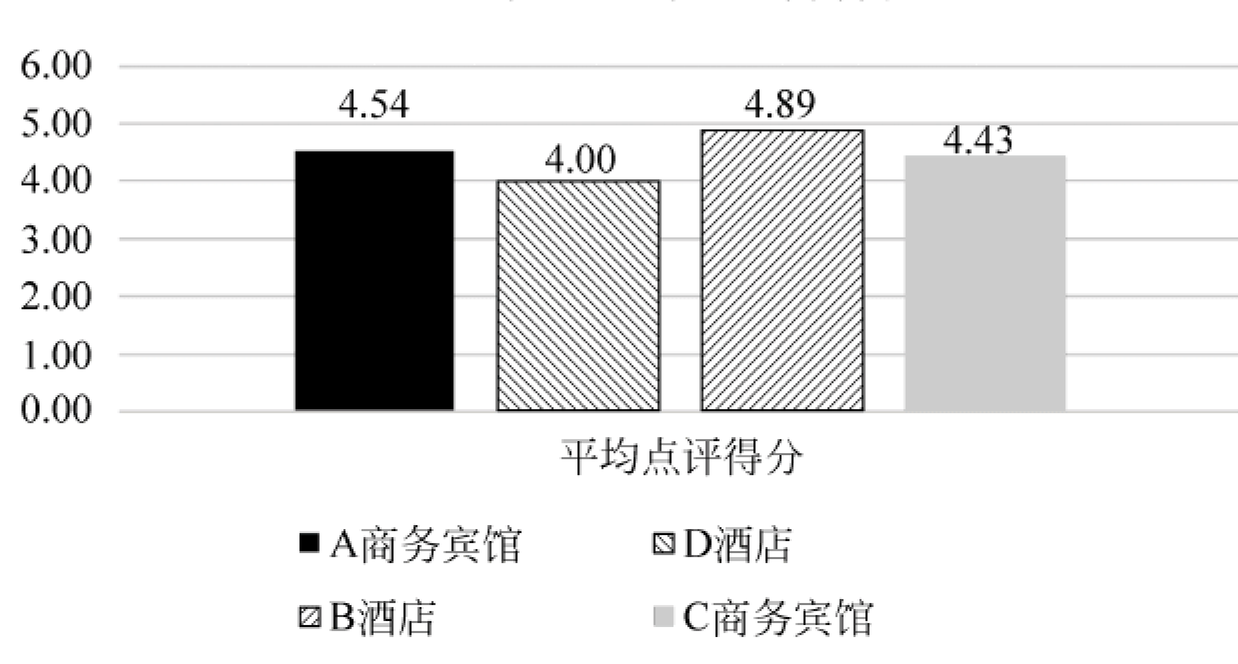


图 8.48 酒店综合评分分布

图 8.48 显示，在平均点评得分上，排名第一的是 B 酒店，其次是 A 商务宾馆，第三名是 C 商务宾馆，最后是 D 商务酒店。B 酒店定位偏高端，在评分上远超过其余 3 家。A 商务宾馆在整体上处于第二的位置，说明其具备一定的竞争优势。

3. 分项评分比较

接下来对 4 项基本点评得分进行讨论分析，首先给出地理位置这一项的平均得分，如图 8.49 所示。

可以看到，地理位置评分分布与整体综合评分分布排序一致，但是值得注意的是，A 商务宾馆相对于其余两家商务宾馆的优势显得比较明显。也就是说，在地理位置方面，A 商务宾馆的优势比较明显，仅次于 B 酒店这家星级酒店。

接下来对设施项进行评分分析，结果如图 8.50 所示。

B 酒店作为星级酒店，其在设施上碾压其余 3 家商务宾馆无可置疑。A 商务宾馆依然是其余 3 家商务酒店中得分最高的一家。不过，相对于上述评分，其与 C 商务宾馆的分差被拉小，说明虽然 A 商务宾馆处于领先，不过优势不大，应该注意及时更新维护设备，从而能够获得更好的竞争优势。服务点评平均服务得分如图 8.51 所示。

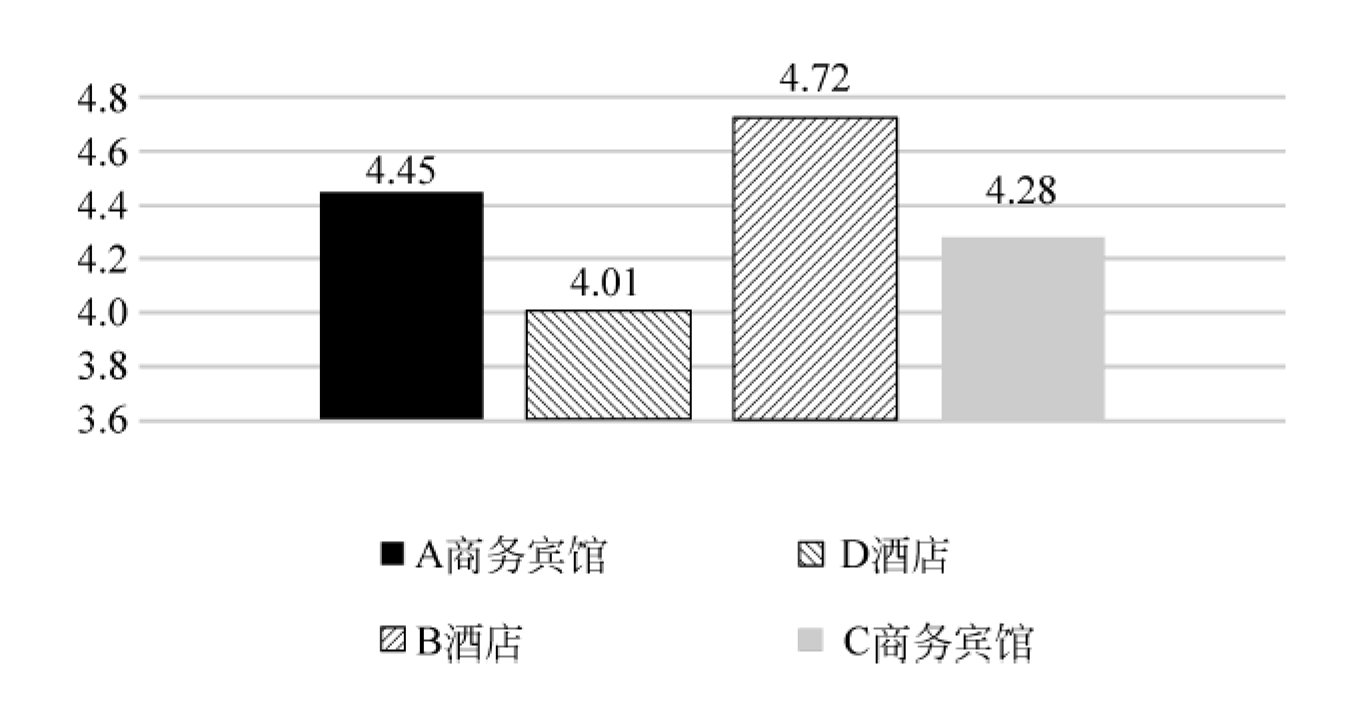


图 8.49　酒店地理位置评分分布

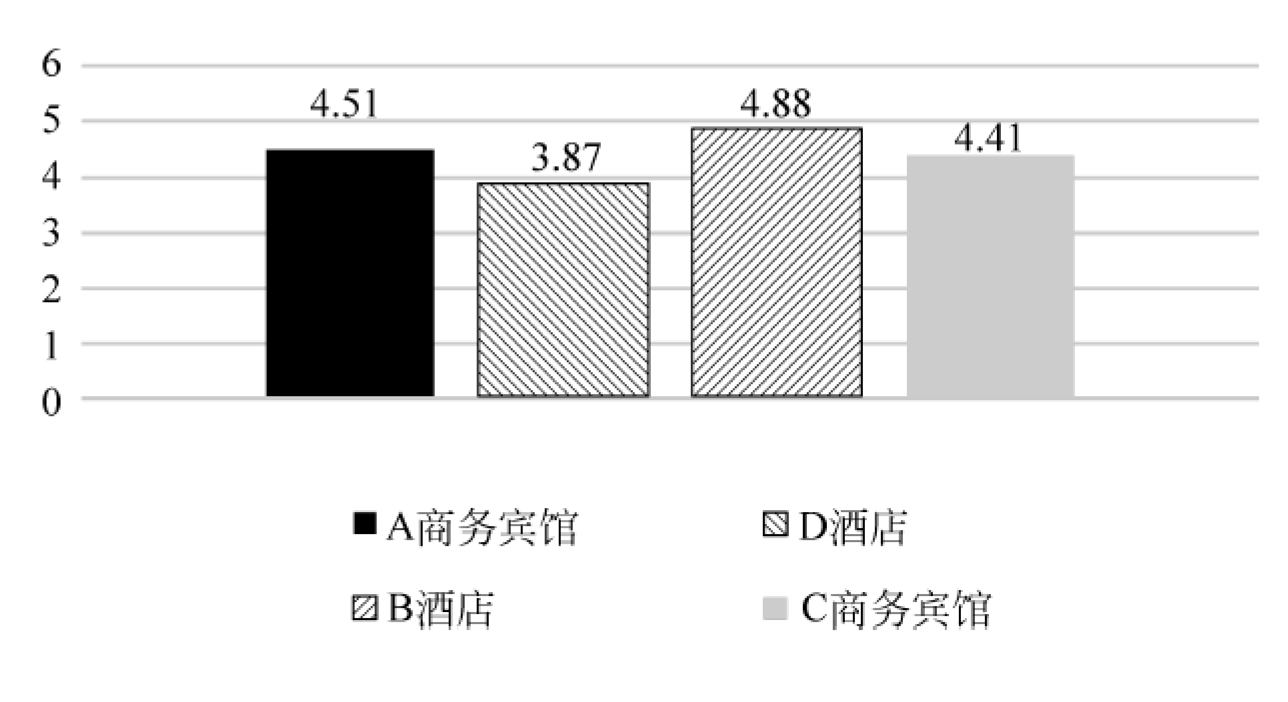


图 8.50　酒店设施评分分布

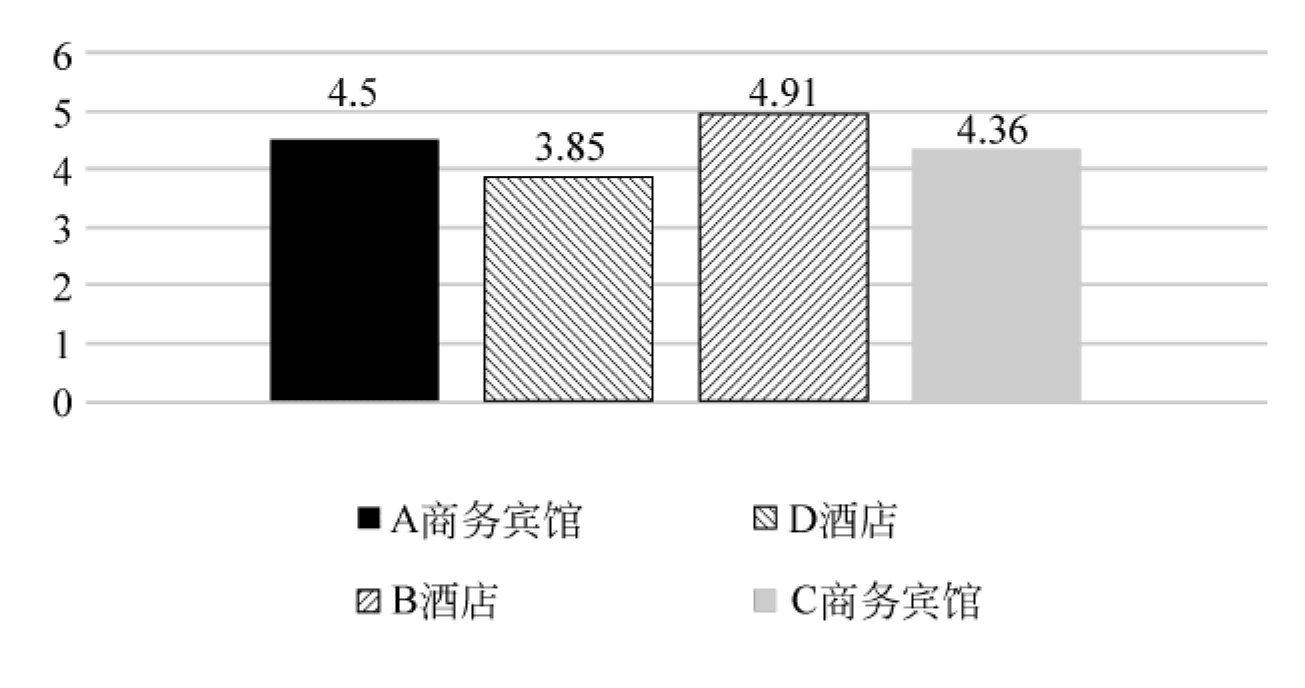


图 8.51　酒店服务评分分布

本项结果中规中矩，排序也与上面两项相同。A 商务宾馆同样相对于其余两家商务宾馆具备竞争优势，但与 C 商务宾馆相差不大，应该及时提升服务质量，从而获得更好的竞争优势。

卫生点评平均得分如图 8.52 所示。本项结果排序也与上面 3 项相同。A 商务宾馆同样相对于其余两家商务宾馆具备竞争优势，不过与 C 商务宾馆相差再次缩小。而 B 酒店 4.93 的分数甩开了其余 3 家商务宾馆很大的距离。对 A 商务宾馆而言，目前竞争最大的是 C 商务宾馆，应该及时提升卫生质量，从而获得更好的竞争优势。

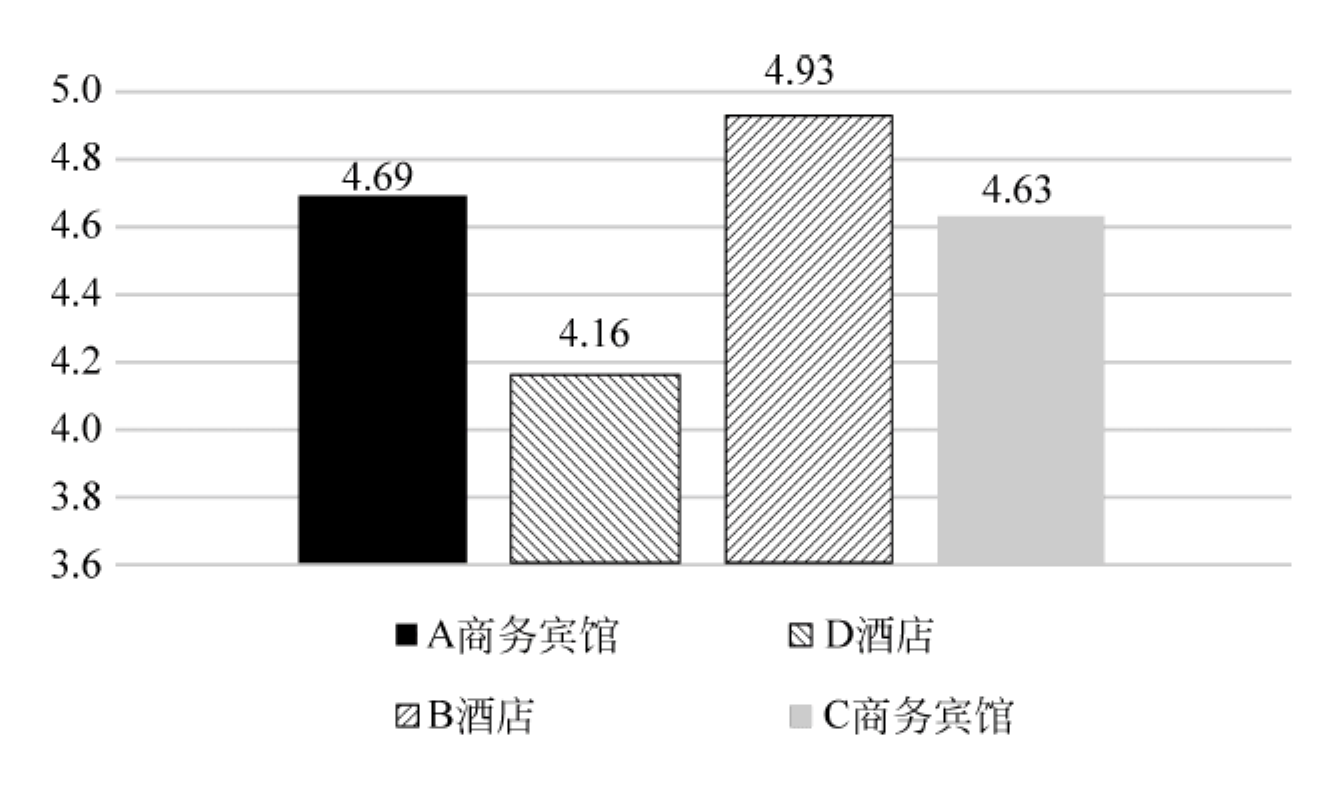


图 8.52 酒店卫生评分分布

4. 客户吸引力对比

在本案例中，某一客人可能在不同的时间入住过不同的酒店，我们将这部分客人的评论数据提取出来，用于对比其对各家商务酒店的评分，并依据入住时间(评论提交时间)对用户的行为进行跟踪，用于对比酒店对客户的吸引力。为了得到更多的客人比较样本，在本节中使用抓取的全部评论数据，即不对评论数据进行随机化删除。

1) A 商务宾馆与 B 酒店比较

A 商务宾馆与周边其他酒店的竞争比较，通过与 B 酒店、C 宾馆对比相同客户对不同酒店的评价实现，见表 8.7。

表 8.7 相同客户对不同酒店评分对比

客户昵称	A 商务宾馆评价时间	A 商务宾馆评分	B 酒店评分时间	B 酒店评分
_CFT010000001287 ****	2016-08-04	4.8	2016-09-06	5.0
pcly80 ****	2016-11-16	4.0	2016-12-10	5.0
203798 ****	2016-03-27	3.5	2016-10-26	4.8
203798 ****	2016-03-27	3.5	2016-12-07	5.0
320027 ****	2016-09-29	5.0	2016-11-19	5.0
205268 ****	2015-11-21	5.0	2016-07-10	5.0
205268 ****	2015-11-17	5.0	2016-07-10	5.0
118843 ****	2016-10-07	4.3	2016-10-27	5.0
118843 ****	2016-10-03	4.0	2016-10-27	5.0
231218 ****	2016-02-16	5.0	2016-08-17	5.0
203798 ****	2016-03-27	3.5	2016-10-26	4.8
203798 ****	2016-03-27	3.5	2016-09-23	5.0
203798 ****	2016-03-27	3.5	2016-12-07	5.0
203798 ****	2016-03-27	3.5	2016-12-07	5.0
203798 ****	2016-03-27	3.5	2016-09-23	5.0

从表 8.7 和图 8.53 中可以看到 A 商务宾馆的用户大部分都是入住之后评分不高，然后流失到 B 酒店的，说明在与 B 酒店的竞争中，A 商务宾馆竞争力较差，并且查看到评分的

时间先后顺序，都是在先住了 A 商务宾馆之后，发现服务各方面不满意之后才选择 B 酒店，虽然两者的客户群体定位不同，但是因为酒店中房型有交叉，即客户群体大致相同，综合说明 A 商务宾馆在与 B 酒店的竞争中处于明显劣势，并存在较高的客户流失风险。

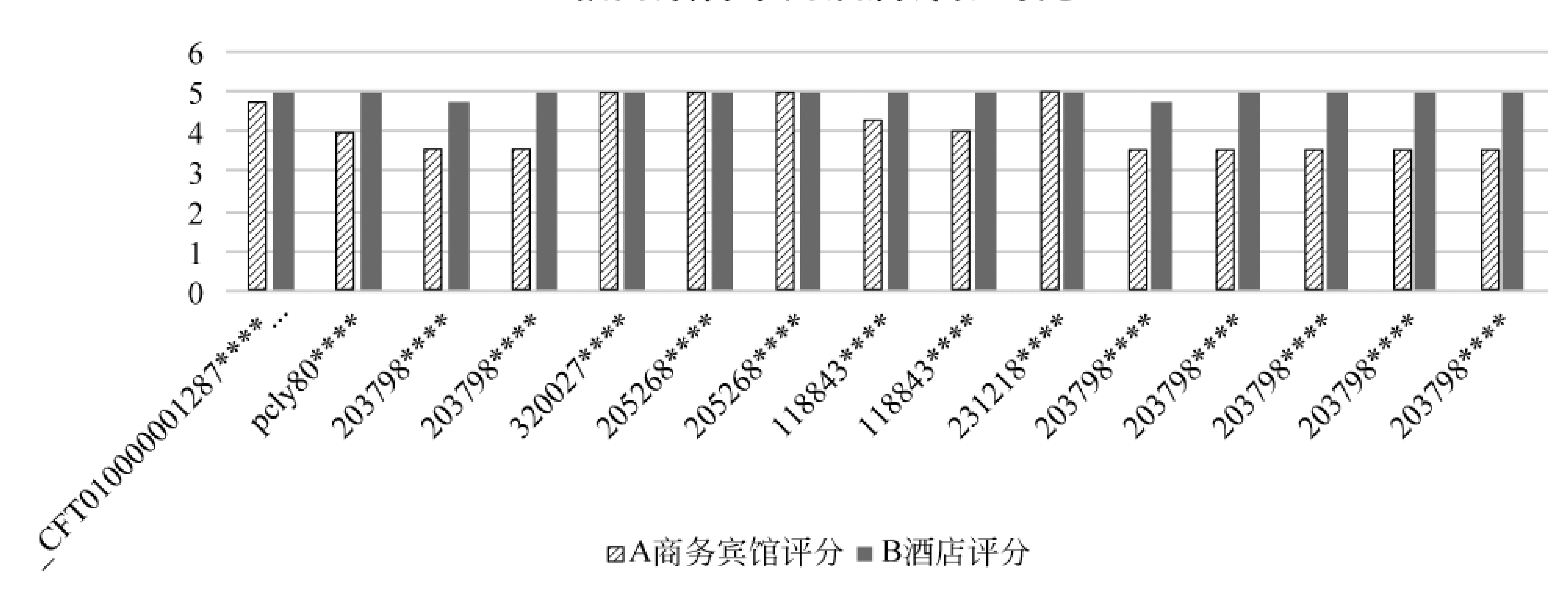


图 8.53　相同客人给 A 商务宾馆与 B 酒店评分比较

图 8.53 中除了少数客人的评分同为 5 分之外，其他客人的评分中全部选择 B 酒店为高评分，并且分差较大，从目前抓取的评论数据中看，此类客人数较少，只占总评论数的 2.2%，但在这部分客人中，除了 4 人给出相同的 5.0 分外，其他的人 100%选择了 B 酒店，可能是双方客人存在差异化，或者目前并未进入直接竞争阶段，虽然如此，A 酒店仍需要提前规划，进行风险防范。

2）A 商务宾馆与 C 宾馆比较

C 商务宾馆在 3 家商务经济型酒店中评价相对较高，以其作为代表与 A 商务宾馆进行比较。表 8.8 是 A 商务宾馆与 C 商务宾馆评分对比，从中可以看到客人同时在两家酒店都有消费，时间点也较多，说明两家酒店的客户重叠率较高，是直接竞争的关系。

表 8.8　A 商务宾馆与 C 商务宾馆评分对比

昵　　称	A 商务宾馆评分时间	A 商务宾馆评分	C 商务宾馆评分时间	C 商务宾馆评分
coolszy	2016-10-27	3.5	2016-09-25	5.0
WZHuangHe	2016-05-08	3.5	2016-11-10	4.3
M13388 ****	2016-08-18	5.0	2016-01-05	5.0
fengji ****	2016-10-12	4.0	2016-07-14	4.8
118002 ****	2016-12-03	4.3	2016-06-17	4.5
品味人生	2016-11-13	5.0	2016-07-31	5.0
品味人生	2016-10-26	5.0	2016-07-31	5.0
品味人生	2016-09-19	5.0	2016-07-31	5.0
品味人生	2016-10-24	5.0	2016-07-31	5.0
jiao_qiao	2016-04-01	5.0	2016-08-24	5.0
y6080	2016-07-12	4.0	2016-02-01	5.0
6851 ****	2016-09-28	5.0	2016-09-26	5.0
M26855 ****	2016-10-30	2.8	2016-10-30	3.0

续表

昵　　称	A商务宾馆评分时间	A商务宾馆评分	C商务宾馆评分时间	C商务宾馆评分
118002 ****	2016-12-03	4.3	2016-06-01	4.0
300489 ****	2016-10-23	4.0	2015-05-09	4.0
300489 ****	2016-10-23	5.0	2015-05-09	4.0
300251 ****	2016-10-23	5.0	2016-10-17	5.0
300251 ****	2016-10-21	5.0	2016-10-17	5.0
300251 ****	2016-10-17	5.0	2016-10-17	5.0
300251 ****	2016-10-18	5.0	2016-10-17	5.0
jiao_qiao	2016-04-01	5.0	2016-07-08	4.0
_M1381895 ****	2016-11-28	5.0	2015-05-13	4.0
_M1381895 ****	2016-06-23	4.0	2015-05-13	4.0
折腾 000	2016-08-24	5.0	2015-03-14	4.0
M10537 ****	2016-07-02	5.0	2015-11-12	3.5
yuxun5200	2016-06-16	5.0	2015-05-28	5.0
yuxun5200	2016-06-16	5.0	2015-05-28	5.0
1590520 ****	2016-05-19	4.0	2015-12-22	3.5
zhangji ****	2016-01-02	2.0	2015-11-28	3.3
zhangji ****	2015-11-28	5.0	2015-11-28	3.3
6851 ****	2016-09-28	5.0	2015-10-21	5.0
6851 ****	2016-09-28	5.0	2015-04-05	5.0
300251 ****	2016-10-23	5.0	2015-08-14	3.8
300251 ****	2016-10-21	5.0	2015-08-14	3.8
300251 ****	2016-10-17	5.0	2015-08-14	3.8
300251 ****	2016-10-18	5.0	2015-08-14	3.8
WZHuangHe	2016-05-08	3.5	2014-11-21	3.8
zhangji ****	2016-01-02	2.0	2015-11-14	4.0
zhangji ****	2015-11-28	5.0	2015-11-14	4.0
zhangji ****	2016-01-02	2.0	2015-10-21	4.0
zhangji ****	2015-11-28	5.0	2015-10-21	4.0

图8.54是其比较结果的直观显示，横坐标为客人昵称，先后顺序代表了时间的前后顺序，对比发现前期A商务宾馆的评分较低，随着时间的推移，有更多的客人从C商务宾馆转向A商务宾馆，说明A商务宾馆在与C商务宾馆的竞争中，客户吸引力有逐步增强的趋势。

相同客人对A商务宾馆打分超过C商务宾馆的有14条记录，低于其分值的记录数为10条，具有71.4%的竞争优势，优势未超过80%，并不明显。从客户吸引力的角度来看，A商务宾馆在同类商务酒店中具有微弱领先的客户吸引力，但在与星级酒店的竞争中明显处于劣势，由于星级酒店B与A商务宾馆的平均价格差低于100元，一旦对方进行促销或推出特惠等降价营销方案，将与A商务宾馆形成直接竞争，会严重影响A商务宾馆的经营。

5. 热词频率对比

通过对评论中热词的分析，我们知道具有较高区分度的核心热词有：服务、早餐、干净、卫生、前台、环境、热情、高铁，其中“早餐”为负面关键词，出现的频次越多，说明评价越低，即

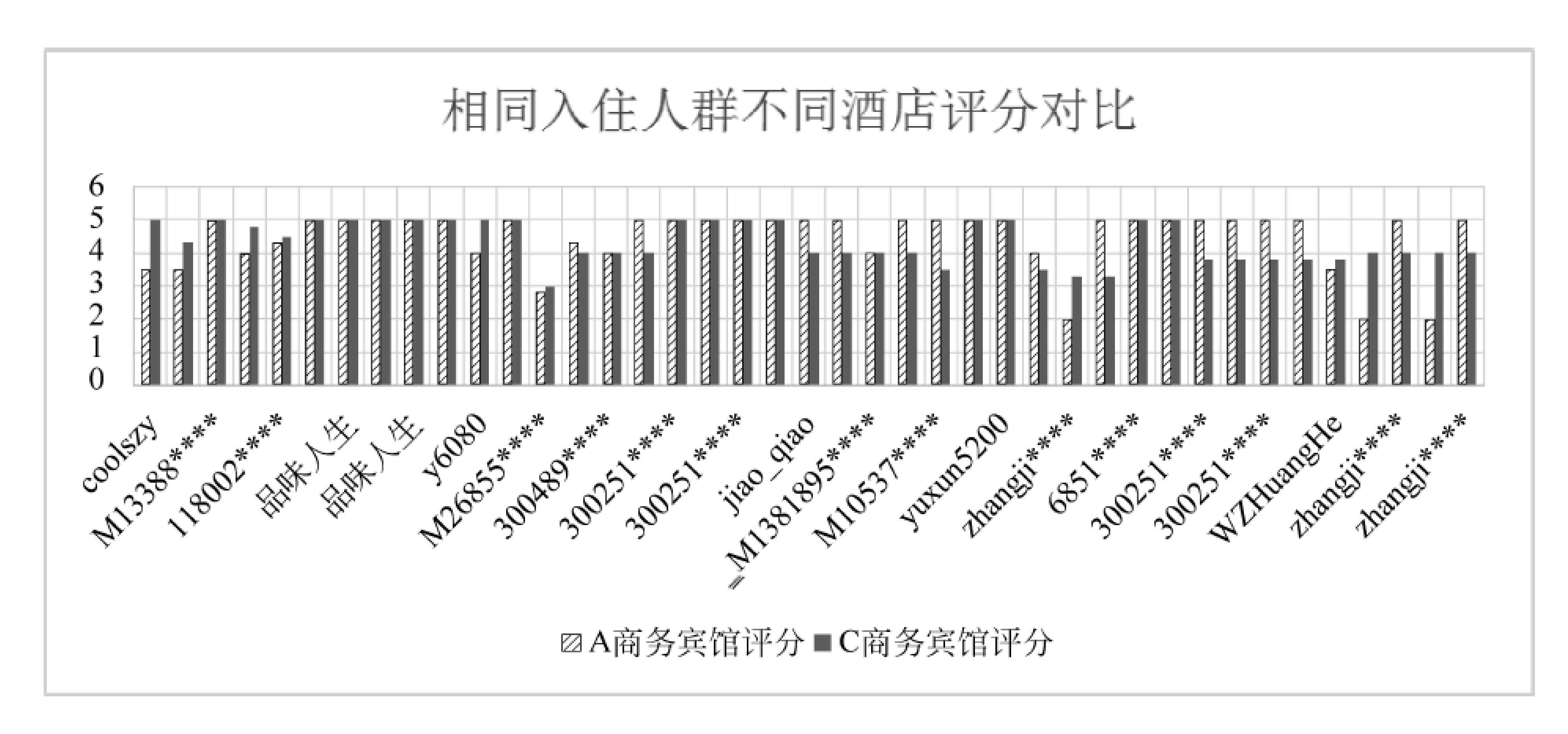


图 8.54　A 商务宾馆与 C 商务宾馆评分对比

一般以抱怨早餐品质为主，而"高铁"这一热词的区分度较低。

由于各家酒店的评论数量相差较多，为了获得各家酒店中各关键词公平的比较结果，使用关键词出现的频率进行对比分析，关键词频率为其词频除以某一酒店的所有分词的总词频数之和，如果某一关键词的比值较高，说明这家酒店的此项特征较明显，如果关键词为正向态度，则表示酒店在这方面较好，反之，说明酒店的问题较严重。经过统计计算，3 家酒店的关键词出现频率对比结果如图 8.55 所示。

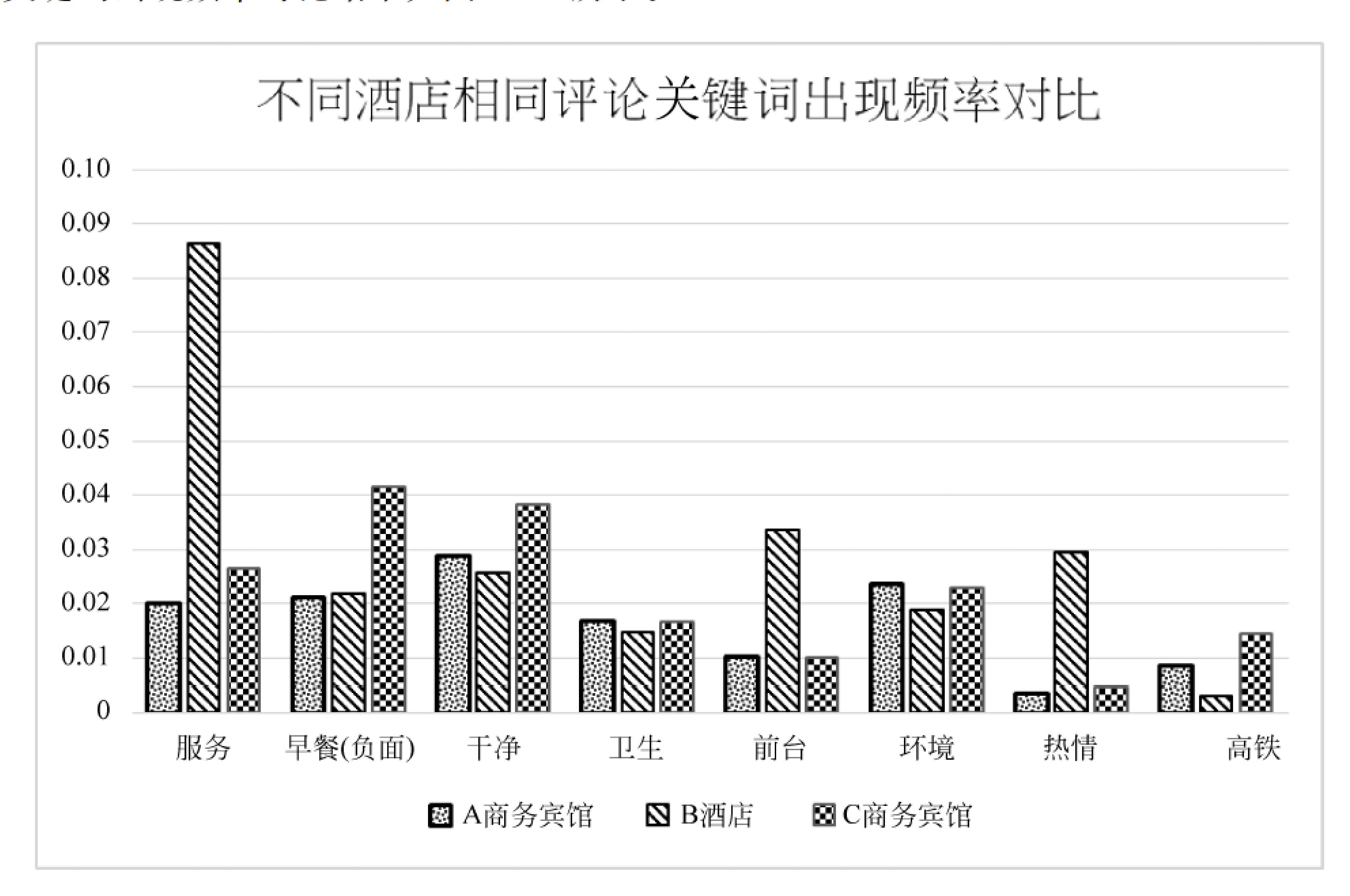


图 8.55　不同酒店相对评论关键词频率对比

从图 8.55 中可以看出，服务方面 B 酒店远高于其他两家商务酒店，A 商务宾馆的服务处于最低水平，略低于 C 商务宾馆。从前文的分析中已经知道，服务水平在酒店的整体评价中最重要，区分度最高，所以在这方面 A 商务宾馆提升空间很大。

“早餐”为负面关键词，C 商务宾馆的评论中提及最多，远高于其他两家，说明这家酒店的早餐质量确实较差，而另外两家酒店的出现频率相差不多，没有太多差距。

在干净卫生和环境方面，3 家酒店的评论中出现频率区分度不大，只是 C 商务宾馆在“干净”关键词表现略好于其他两家酒店。

提及“前台”的评论中，B 酒店中的频率远高于其他两家酒店，同时出现“热情”的评论，B 酒店更是远超过它们，结合服务方面的表现，说明 B 酒店的服务水平有口皆碑，已经形成较强的品牌影响力和竞争力，A 商务宾馆和 C 商务宾馆之间的频率结果相差不大，间接说明经济型商务酒店在服务方面，具体表现就是前台不热情或者其热情程度没有给客人留下较深印象。

“高铁”的关键词区分度一般，其结果的准确性较低，C 商务宾馆和 A 商务宾馆都位于高铁站附近，所以这两家的评论中提及频率较高，这并不影响酒店的整体评价水平，也间接说明酒店的位置对酒店的好评率影响较小。

6. 情感分析对比

基于前述对客户情感的分析理论，依据爬虫获得评论数据，对评论内容进行情感分析，得到不同酒店之间的情感分布图，如图 8.56 所示。

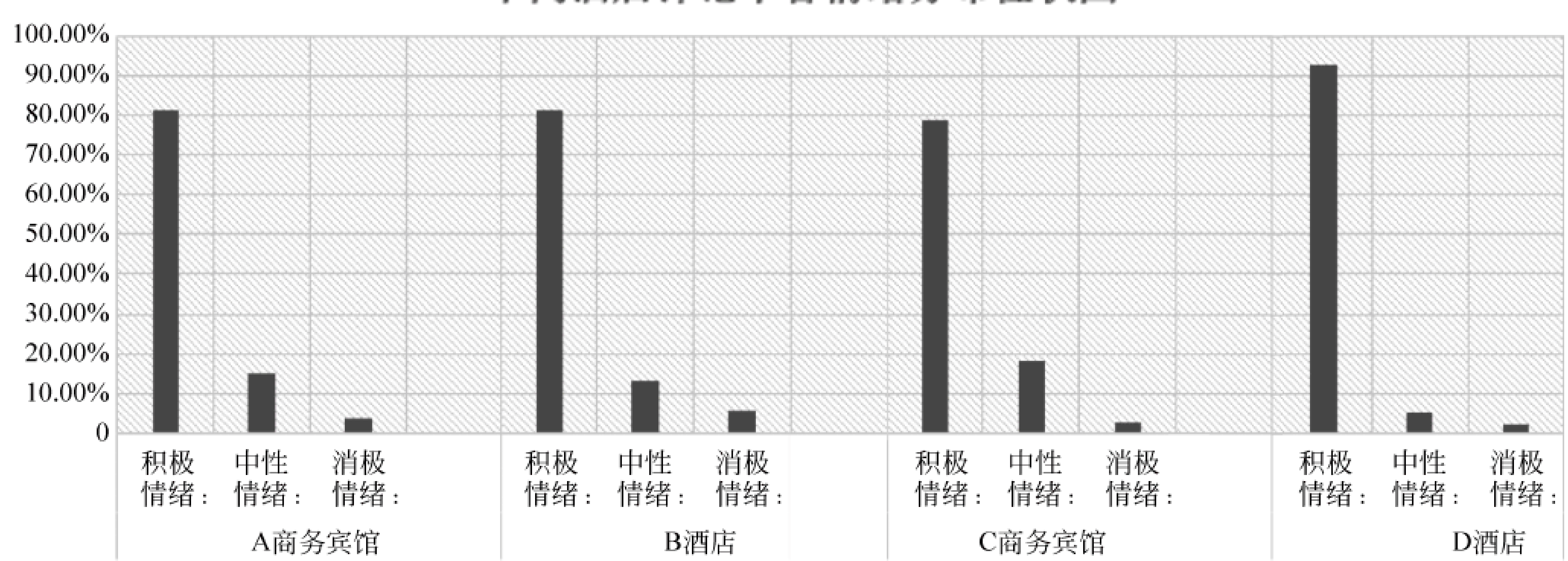


图 8.56 不同酒店情绪分布图

从图 8.56 中可以看出，在评论积极情绪方面，好评率最高的是 B 酒店，其次是 A 商务宾馆，然后是 C 商务宾馆和 D 酒店。从中可以看出，对于 A 商务宾馆而言，在业界好评最高的是 B 酒店，也是 A 商务宾馆强有力的竞争对手。将 A 商务宾馆、C 商务宾馆、B 酒店的情感分词的标签云进行对比分析，如图 8.57 所示。

从评论分布中可以清晰地看出，B 酒店的服务明显更为高档，如还提供水果、自助早餐等服务，可以看出在服务方面很热情，其中前台的作用非常重要。这与 B 酒店的自身定位密切相关，其作为一家连锁星级酒店，在好评的积极情绪上比 A 商务宾馆更佳。

从图 8.57 中可以看出，A 商务宾馆和 C 商务宾馆的最高词频都是方便、干净、环境、早餐等。观察词频可以发现 B 酒店的“服务”非常突出，词频为 632 次，“前台”被提及的次数有 250 次，这是其主要优点之一，说明从事服务行业前台的服务水平非常重要。

图 8.58 是根据负面评论内容进行分词生成的可视化词频统计。其中，A 商务宾馆被用



图 8.57　从左至右依次为 A、C、B 酒店全部评价主题

户提及最多的是早餐、态度、难找等；C 商务宾馆被提及最多的是早餐、服务、前台、周边等；B 酒店被提及最多的是万达，其次是广场、设施、环境等，之所以万达和广场这两个字在负面词频中较多，是 B 酒店与万达无关，并且距离万达广场并不近，导致客户对此抱怨。



图 8.58　差评标签云图，从左至右依次为 A、C、B 酒店

结果表明，A 商务宾馆令用户最不满意的是早餐问题，词频为 4 次，服务员的服务态度也不够到位，很多客人提及位置难找和周边配套，并且存在房间装修异味问题。C 商务宾馆同样存在服务问题，存在的主要负面关键词与 A 商务宾馆基本一致，早餐情况要差于 A 商务宾馆，词频为 12 次，从侧面可以看出经济型商务酒店在服务和早餐上是主要短板。观察 B 酒店的结果，其负面评论关键词数量明显较少，最高词频仅为 3 次。在全部评论中提及次数最多的服务也出现在差评的高频词汇中，通过查看评论原文发现，是客人点评中首先评价其服务不错，同时批评其他方面不足，导致服务关键词在负面词频中计数多。令用户不满意的是万达广场的关系问题，这是由于 B 酒店距离万达广场超过 1km，并且与万达广场没有关系，除此之外的是有一位客人评价态度较差，一位客人评价晚上停电，一位客人评价环境太偏，一位客人评价网络较慢。总体负面评价较少。

对于定位相同的商务宾馆，可以看到从好评率上 A 商务宾馆具有一定的领先优势，对比三者的评论分布图后发现，A 商务宾馆在地理位置上具有最大优势。不过，同样也可以看出，在服务和设施方面还有待加强。如果能在服务上对宾馆质量进行提升，可以使得 A 商务宾馆远超对手，获得更大的竞争优势。

8.6 建议

经过上述数据挖掘和对各酒店评分及情感分析比较，在A商务宾馆的当前经营竞争条件下，可以得出以下结论：

(1) A商务宾馆在口碑上次于B酒店，略高于C商务宾馆，领先于D商务酒店，在经济型商务酒店中竞争力较强。

(2) A商务宾馆的主要竞争对手是C商务宾馆，两者在综合口碑评分上相差不大，甚至在卫生、设施等单项评分上也难以拉开距离。

(3) A商务宾馆与星级酒店（如B酒店）的竞争中，优势较少，客户吸引力较弱，两者的定位和价格虽有差距，但并不明显，如果B酒店推出低价格客房，将形成直接竞争，最终会影响A商务宾馆的经营收入。

在服务性行业中，A商务宾馆要形成更强的竞争优势，核心在于提高服务水平，不只是在早餐、设施等硬件方面提升客人体验，更要在观念上进行改变，使服务人员发自内心为客人服务，建议在以下几个方面进行改进或完善。

1. 提升服务水平

对员工进行标准化培训，必要时进行商务礼仪培训，提高商务出差客户的服务能力，特别是前台人员的服务意识要加强，使客人能够感受到服务人员的热情，必要时建立前台人员奖励机制，如奖金与网上评论的评分进行挂钩，形成正向激励。

2. 改善早餐品质

早餐是经济型商务酒店的重要竞争因素，除了满足商务人群对早餐的干净、卫生等基本要求外，还需要增加可选种类，以提供多样化选择，适应大多数客人的需求，重点提高餐食的口味品质和服务人员的服务意识。

3. 建立会员系统和关怀方案

对会员投诉进行物质奖励或积分补贴，使客户负面情绪转化为正向好评，可提升客户的忠诚度，提高已入住客户的复购率。

4. 对房型进行重新调整

使客人选择更平均化，不要全集中于舒适大床房和舒适双床房。目前房型的区分度只在于面积大小，并不实用，建议在高端房型中增加餐饮方面更加优质的服务，如提供更多早餐选择、入住即送免费水果、下午茶、夜宵等，特别注意适应商务人群的高层次需要。

5. 使客户易于找到酒店

尽可能在主要道路建立指引标识；在酒店预订网站的相关网页介绍中加入指引提示。

6. 保证设施和硬件维护及时

及时更新或优化设施，如空调、Wi-Fi设备、地毯、热水等，建立应急处理机制，保证客人报修后及时响应，防止出现维修不及时，给客人带来不便。

第 9 章

耐热导线工厂质量管理数据分析

随着制造企业信息化的发展，生产过程逐步实现数字化，企业会积累大量的制造和质量检测数据。在大数据时代，如何利用这些数据，从中找出产品生产过程中存在的问题，发现制造流程中可以改进的环节，这是减少制造成本，提高产品质量的重要保证，也是实现智慧工厂的必要组成部分。

9.1 项目概述

某集团耐热导线工厂(以下简称耐热导线工厂)在多年的生产过程中，已经上线了基本的生产管理系统，收集了产品生产过程中的一些工艺参数、各工序的成品检测结果等数据，通过对这些数据的分析，可以在很大程度上减少经验式管理带来的不足，降低废品率，提高加工机台的工作性能和稳定性。

目前，耐热导线工厂的主要产品包括钢芯铝绞线、钢芯铝合金绞线、铝合金绞线以及铝包钢绞线等系列产品，并致力于高强度铝合金线、耐热铝合金导线等新产品的开发生产。目前，多种产品在国内外的市场上占有一定的地位。

耐热导线的生产主要由三道工序组成：轧机、拉丝和绞线，不同工序对质量都有相应的要求。轧机工序的成品质量与后续的两个工序的成品质量之间有明显的关联性，因此可以通过第一道工序的成品质量预测后续工序的成品质量，也可以用后来两道工序产品的质量来“反推”第一道工序的质量。

目前，公司对于耐热导线的制造数据管理还停留在检测数据的简单录入、查询阶段，有关机台工艺参数和加工状态的数据，还暂时没有收集或充分利用，难以通过数据分析技术建立工序间的关联性，因此不容易在生产前进行预警、在生产中进行控制，往往到最终的产品检验时发现质量问题为时已晚。我们在耐热导线工厂最近 2 年的质量管理数据的基础上，分析了这些数据存在的问题，进行了大量的预处理，利用统计学、多维度分析、数据挖掘以及可视化等多种数据分析方法，以提高最终产品的合格率为目标，探索耐热导线的加工流程中几个步骤之间半成品或成品质量指标之间的关系。

影响耐热导线加工过程的因素很多,这些因素或多或少还存在一定的相关性,它们之间的关系使用数学函数表示,也是非常复杂的非线性函数。但分析不同工序质量指标的相关性,在很大程度上可以调节不同工序的加工要求,使最终产品的合格率提高。此外,有一部分成品的质量规格超过的国家标准较多,这说明第一道工序中存在着“质量冗余”,需要通过分析工序之间的关系,在产品合格的基础上使质量指标达到一个较合适的标准,以消除冗余,降低成本。

限于篇幅以及保密原因,本章仅讨论单线线径与所选用杆强度之间的关系,即是否有必要按照单线线径的范围来选择相应的杆强度范围。

9.2 耐热导线生产质量数据预处理

我们曾多次与耐热导线工厂相关人员沟通,并赴工厂实地考察耐热导线生产过程。耐热导线生产的原材料是铝杆,公司接到订单,确定生产某种具体规格的铝线后,根据杆材流转使用规定选择相应的铝杆,并检查铝杆是否符合相应的要求。然后进入铝线的生产工序——拉丝工序。经过高速拉丝后,通过检验铝线的线径、表面质量等指标,对铝线的质量进行控制。

数据来源为耐热导线工厂提供的自 2014 年 3 月开始至 2016 年 2 月底两年的铝线生产线生产数据,包括原材料检测数据 50 万条,成品检测数据 70 万条,制造执行系统(MES)中各条生产线的制造数据总计 150 多万条。直接对如此多的数据进行处理和分析难以满足要求,因此需要对工厂提供的原始数据进行整合和预处理。

耐热导线工厂提供的数据来自于原来的项目执行单表、轧机生产日报表、拉丝生产日报表、绞线生产日报表、各类成品检测表、各类半成品检测表、各类原材料检测表、机台设备信息表、班组信息表等。数据量大且较为分散,需要的信息分散在多个数据表里。为了根据目标铝线选择相应的铝杆,就需要通过耐热铝线的编号追溯到铝杆的各项数据。通过将订单编号与项目执行单进行关联,项目执行单与轧机生产日报表、拉丝生产日报表、绞线生产日报表通过相同合同编号进行连接,进行关联的方式,追溯铝杆的生产数据。

这里主要采用 SQL Server 中的 T-SQL 语句 inner join、left join、right join 等将多表进行连接整合,得到与铝杆相关的数据主要包括铝杆的重量、实测外径、抗拉强度、拉断力、伸长率、正向/反向电阻值、20℃时电阻率、室温以及与铝线相关的主要参数铝线的线径与抗拉强度等。整合后得到目标铝杆参数、铝线参数表。

数据分析能获得数据中蕴藏的信息或知识。高质量的数据是数据分析的基础。我们在耐热导线的数据分析过程中,主要使用了导线加工过程各工序的质量检测数据,而加工设备、生产工艺以及人员的数据因为保密、数据收集不全等原因暂时没有使用。我们也发现了耐热导线生产过程中数据收集的一些不足,如有些数据人工输入错误或者测量有误差,某些有用的数据暂时没有收集或缺失,这些问题都对耐热导线数据分析的结果产生了一定的影响。

耐热导线工厂提供了近两年耐热导线检测的数据,涉及多个合同、多个批次以及多个加工机台。经过上述数据整理的步骤之后,数据中还存在着“脏数据”。所谓脏数据,就是数据中存在噪声数据、错误数据、缺失数据以及冗余数据等问题。数据清理在数据预处理阶段花

费时间占比最大，但同时它也是最重要的步骤，该步骤可以有效减少脏数据造成的低质量分析结果。

1. 噪声数据处理

噪声数据是指数据中存在着错误或偏离期望值的数据，引起噪声数据的原因可能是硬件故障、编程错误、拼写错误或者识别程序中的乱码。对于噪声数据，尤其是孤立点或异常数据，不能随便删除，这些数据很可能是数据分析中的异常数据。

在耐热导线生产数据（这里主要是各工序成品质量检测数据）中，主要存在的数据噪声形式有缺失值、异常值、冗余值等。缺失值主要存在生产过程记录的数据中，并不是每一项指标的数据都有记录，同一批铝杆拉出的铝线，时常会出现某个铝线线径空缺的情况，这主要与工厂的生产线记录管理有关。异常值是指存在一些明显不符合常规的数据，如有些记录中铝线的线径数值达到了423mm，而这个数实际可能取值是4.23mm，属于手工录入错误。冗余主要表现为同一生产批次的数据重复出现，这往往是数据整合过程中产生的问题。

在考察铝杆抗拉强度与铝线线径关系时，首先采用分箱技术。由于耐热导线工厂给出的数据中，同一个抗拉强度对应的线径有时差距非常大，存在一定的噪声数据，根据抗拉强度对数据进行分箱处理。然后对同一个箱子里的数据进行处理，将数据样本中的奇异值、极端值、非正常值等数据以及数据本身的特点采用图形方式呈现出来，并剔除非正常的数据样本。反映变量集中趋势的有算术平均数、中位数。反映变量离散程度的有方差、标准差和极差。反映分布形态的描述性指标有偏度（skewness）和峰度（kurtosis）。偏度和峰度是判断数据是否正态分布的重要指标。在实际检验中，偏度和峰度都小于1时，可以认为数据近似服从正态分布。

1）按照拉依达准则（3σ 准则）剔除异常值

拉依达准则是在数据总体服从正态分布的情况下，根据下面公式找出异常值：

$$p(|x-u|>3\sigma)\leqslant 0.003$$

式中，u 表示变量的平均值；σ 表示变量的标准差。对大于 $u+3\sigma$ 或小于 $u-3\sigma$ 的数据作为异常数据，予以剔除。剔除后，对余下的各测量值重新计算偏差和标准偏差，并继续审查，直到各个偏差均小于 3σ 为止。例如，在处理某批抗拉强度为123MPa的铝杆对应的铝线数据时，根据描述统计得到其对应的铝线线径分布近似正态分布，可以运用拉依达准则将异常值剔除。

2）按照时间序列平滑数据

考虑到在实际测量过程中铝线线径数据可能出现的测量误差，在利用拉依达准则剔除异常值后，利用多次测量取平均值的误差消去方法，对一个箱子中时间间隔在3min内的铝线线径数据取平均值平滑处理。

2. 缺失值处理

处理缺失数据的方法有多种：可以采用近阶段数据的线性插值法进行补缺；可以采用该时间段的历史数据填补丢失时间；可以用缺失数据样本周围的数据来代替；可以使用一个全局常量或者属性的平均值填充空缺值；可以使用数据挖掘的算法对数据进行修复，如回归方法、决策树或者贝叶斯方法；也可以删除少量的空值。

由于铝线与铝杆的检测参数均为连续性的数值，而且同一生产批次产品的检测参数都在一定的小范围内变化，所以主要采用线性插值法对缺失参数进行补缺。

3. 冗余数据处理

在数据整合阶段，将数据由不同的业务表整合在一起，有些记录会产生重复的情况。例如，将生产日报通过订单号进行关联时，可能会产生多条相同订单号的生产记录。因此需要对多条相同订单号的生产记录数据进行处理，通过 SQL 中的 distinct 关键字对冗余的数据进行过滤，只保留一条数据。

4. 铝线生产数据的归约

由于存在多个铝杆属性，这些变量之间可能存在某种关系，会导致变量在表达某一现象时产生重叠性。数据预处理时，首先得到对输出变量影响较大的输入变量，保留这些变量并剔除分析后明显不相关的变量，约简变量个数。考察相关性时，本项目采用皮尔森(Pearson)相关系数和决策树分析两种方法。

1) 皮尔森相关系数法

皮尔森相关系数是用来反映两个变量相关程度的统计量。当两个变量的线性关系增强时，相关系数趋于 1 或－1。正相关时趋于 1，负相关时趋于－1。当两个变量独立时，相关系数为 0。采用皮尔森系数考察变量之间的相关性。

可以看出，与铝线线径相关性较高的两个属性分别是铝杆的抗拉强度、铝杆的伸长率，而其他变量与铝线线径相关性的绝对值都小于 0.1，且显著性水平大于 0.05。例如，铝杆的电阻率与铝线线径的皮尔森相关系数为－0.008，而且显著性检验为 0.767，杆材电阻率和线径无关的概率到了 0.767，显然无法拒绝该假设。而杆材的伸长率与铝线线径的相关系数达到了－0.415，显著系数为 0.000 表示杆材伸长率与铝线实测外径无关成立的概率为 0，可以拒绝该假设，证明铝杆抗拉强度与铝线线径存在强相关性。

值得注意的是，伸长率与抗拉强度的负相关性也达到了 0.651，如果考虑两个变量对铝线线径的回归分析，会产生共线性的问题。这里的共线性问题是指回归模型中的自变量之间由于存在相关关系或高度相关而使模型估计失真，或难以估计准确，需要使用特殊的回归模型处理。

2) 决策树分析法

为了进一步验证通过皮尔森系数法得到的相关关系，本项目还采用决策树模型对铝线线径的影响因素进行分析。这里采用基于 CART 算法的决策树考察影响铝线线径的影响因素。生成树的过程中使用“剪枝”方法，先建立一个划分较细的树模型，再根据交叉检验(cross validation)的方法估计不同“剪枝”条件下各模型的误差，选择误差最小的树模型。

我们主要关心变量的重要性，可以看到，铝杆的抗拉强度和伸长率居于重要性前两位，而其他变量的重要性不明显，这与通过皮尔森相关系数法得到的结论吻合。因此，在分析过程中，主要考虑伸长率和抗拉强度等属性对铝线线径的影响。但考虑这两个属性带有比较强的相关性，我们只选择铝杆的抗拉强度分析与铝线线径的关系。

9.3 耐热铝线质量检测数据分析

在耐热铝线检测数据中选择了数据较多的 4 个合同。依据单线是否合格，将每个合同下的数据分成两类，然后分别取出合格单线与不合格单线对应的下机数据和冷测数据。再

通过将合格单线对应的下机(或冷测)数据与不合格单线对应的下机(或者冷测)数据进行比较,观察合格与不合格单线对应的两道工序之后的检测数据在性能指标的概率分布异同。

从整体上说,这 4 个合同的主要性能指标(如抗拉强度、电阻率、伸长率等)的概率分布基本相同,说明了规律有一定的通用性,但在细微的地方也有一些差异,体现了每个合同的特殊要求。初步发现的规律如下:

(1) 通过观察单线几个性能指标的概率分布,可以得出这样一个结论:单线不合格绝大多数是因为抗拉强度不合格导致的。因为可以很清楚地看到,不合格单线的抗拉强度几乎全部分布在某一数值左侧,合格单线几乎全部分布在右侧,而且概率分布图中间有断续。而反观两部分单线在其他合同上的分布,它们的均值可能略有差别,但概率分布图几乎完全一致。

(2) 观察单线在抗拉强度上的比较,会发现合格单线对应的抗拉强度均值大于不合格单线对应的抗拉强度均值,并且合格单线对应的冷测阶段和下机阶段的抗拉强度都相应地大于不合格单线对应的数据,这定性地说明在抗拉强度上冷测阶段和下机阶段的抗拉强度与单线的抗拉强度具有一定的正相关性。但是不合格单线和合格单线对应的冷测和下机抗拉强度均值相差不是很大,而且标准差很大。

(3) 针对某一合同,在抗拉强度这个指标上比较,可以发现合格与不合格单线的抗拉强度均值的差大于对应的冷测抗拉强度均值的差,而后者又大于对应的下机抗拉强度均值的差。这说明冷测对最后性能的影响大于下机对最后结果的影响。

将单线的数据作为因变量输入,冷测的数据作为自变量输入,回归分析得到图 9.1。

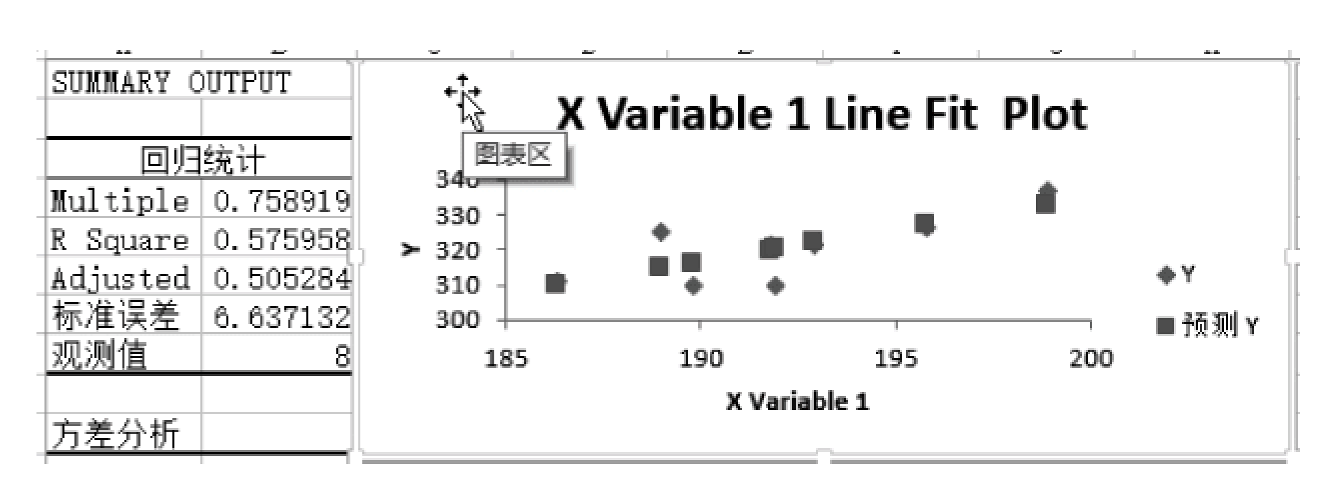


图 9.1　单线与冷测数据的关系

将单线的数据作为因变量输入,下机的数据作为自变量输入,回归分析得到图 9.2。

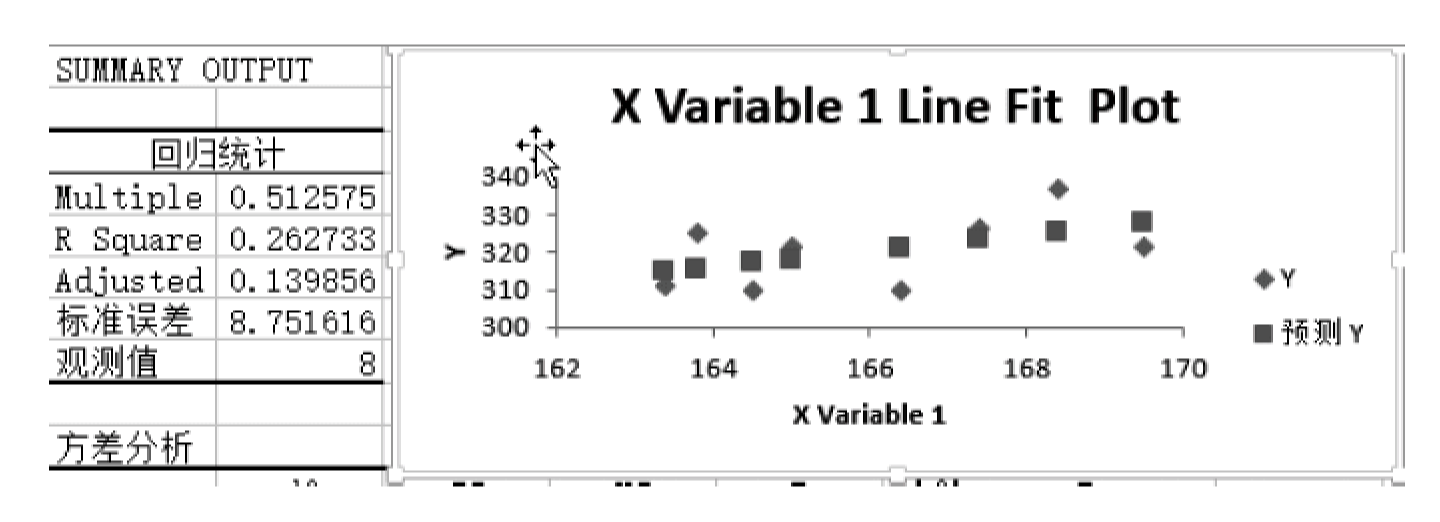


图 9.2　单线与下机数据的关系

可以看到,对于不同合同的数据,这种正相关性都存在。下面以合同号 XX/10789-1 的样本数据为例,对单线抗拉强度进行比较,其中左边为不合格样品的数据,右边为合格样品的数据,如图 9.3 所示。

因为 315MPa 为单线抗拉强度合格最低要求,所以不合格的抗拉强度全部分布在

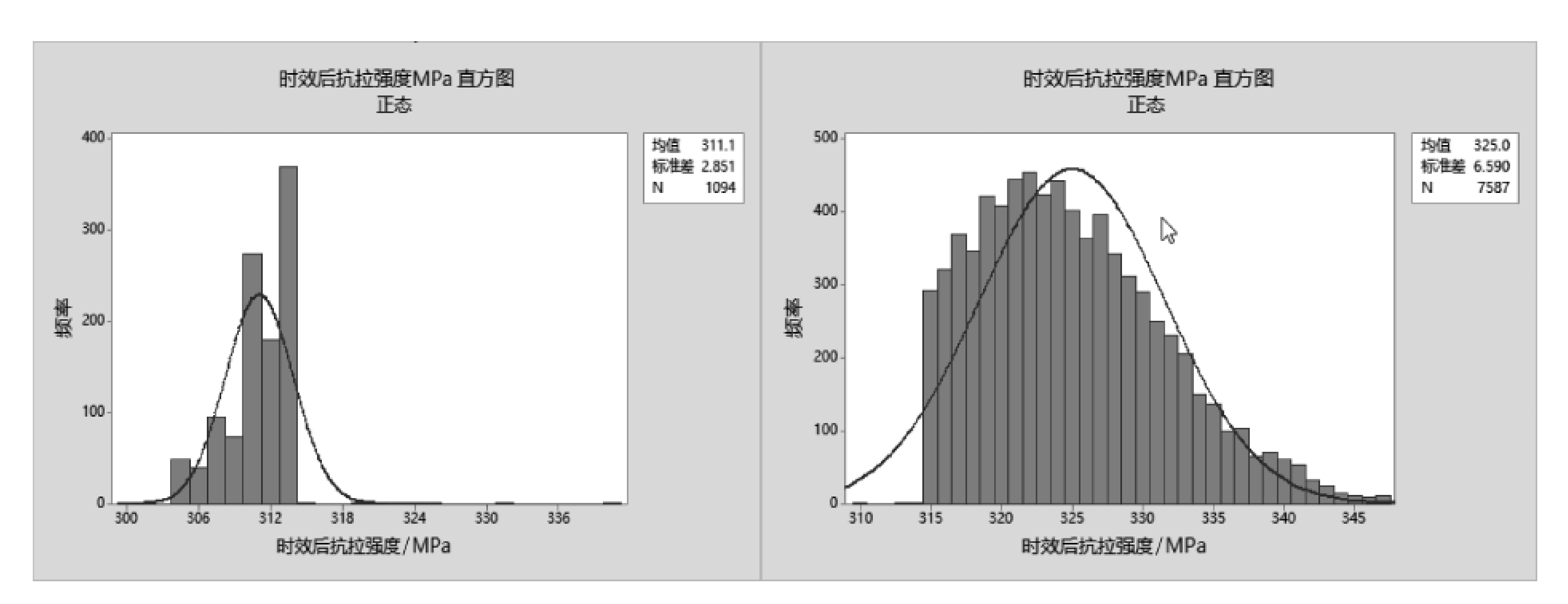


图 9.3 单线时效后抗拉强度的比较

315MPa左边,合格的基本全部分布在315MPa右边,但也有少许单线抗拉强度大于315MPa,但结果为不合格,这可能是由于其他指标(电阻率和伸长率)不合格导致,从数量上也反映出绝大部分不合格成品是因为抗拉强度不合格导致。但也有少许抗拉强度小于315MPa,但记录为合格,这或许是因为员工操作失误导致。类似地,冷测抗拉强度的比较如图9.4所示。

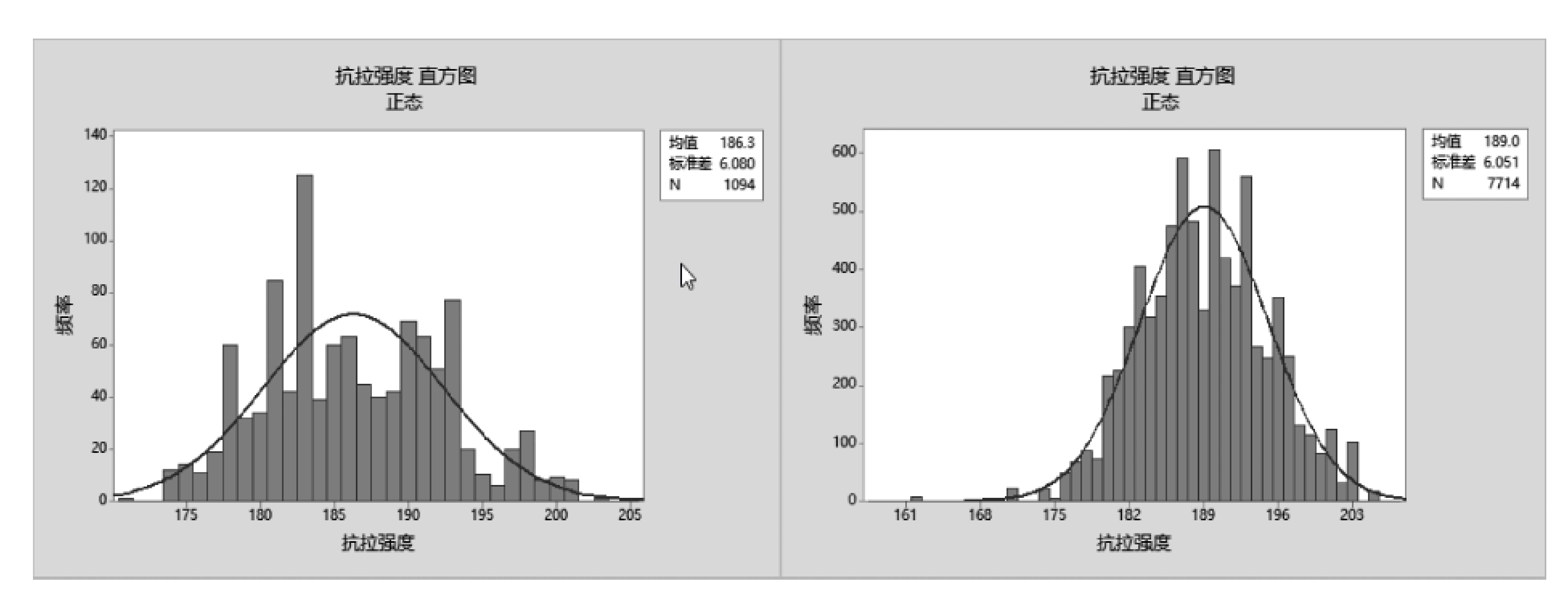


图 9.4 冷测抗拉强度的比较

从图9.4可见,合格和不合格单线冷测抗拉强度的标准差几乎相同,但不合格样品的冷测抗拉强度均值低于合格样品的冷测抗拉强度。图9.5为下机抗拉强度的比较。

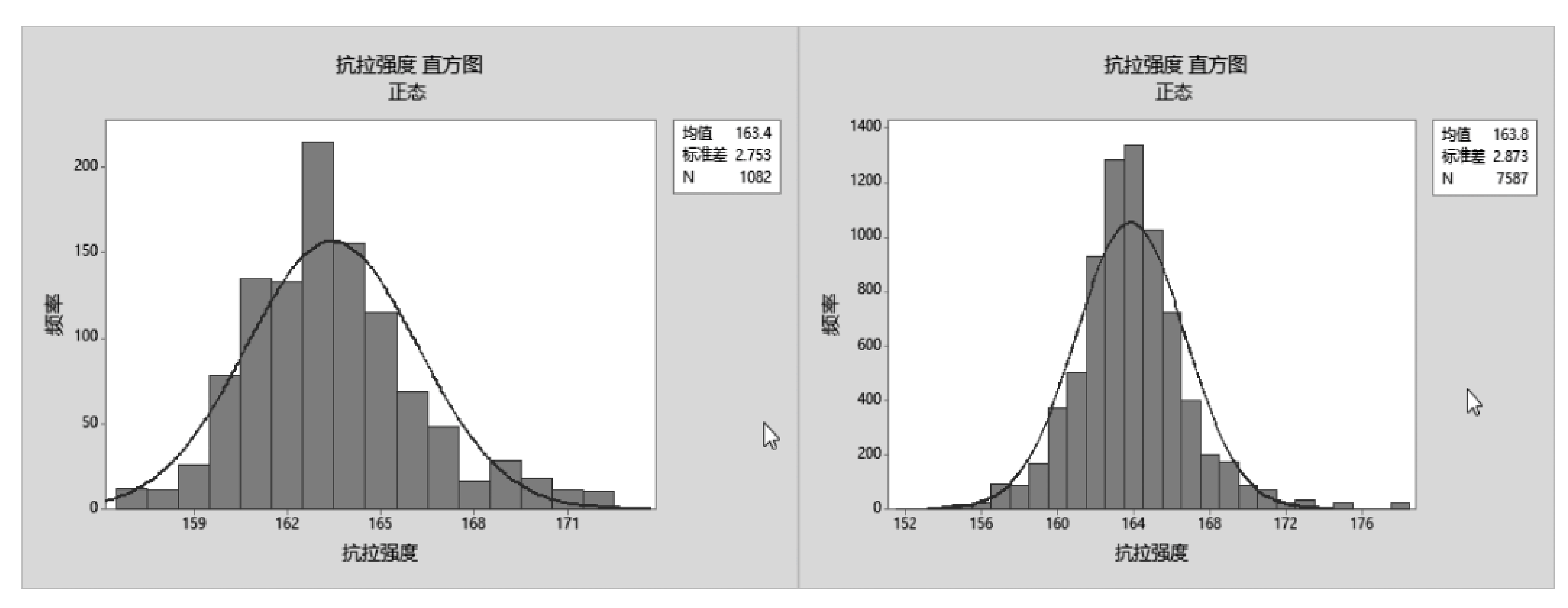


图 9.5 下机抗拉强度的比较

不合格与合格样品对应的下机阶段抗拉强度无论均值，还是标准差都非常接近，几乎没有区别。其他合同号的样品也存在类似的现象。

下面通过铝合金单线线径与选用杆强度的回归分析，分别讨论铝合金与硬铝的单线线径与对应杆抗拉强度之间的关系。这里以 XXX1 型铝合金为例，先给出回归分析得到的结果，见表 9.1。

表 9.1　XXX1 型铝合金单线线径规格与杆材强度范围关系

线径/mm	选用杆强度范围/MPa
2.0～2.5	170.41～187.81
2.5～3.0	175.12～193.56
3.0～3.5	179.83～199.32
3.5～4.0	184.53～205.08
4.0～4.5	189.24～210.84
4.5～5.0	193.95～216.60

XXX1 型铝合金单线线径与所选用杆材强度之间存在较强的线性关系，可以按照单线线径的范围选择相应的杆材强度范围。随着线径变粗，杆材选取的范围会变大。下面给出上述结论的分析过程。

针对上面分析的问题，从数据集中选取 3 个属性变量，分别为“铝合金类型”“杆材强度”以及“线径规格”。其中，“杆材强度”与“线径规格”是分析的对象，“铝合金类型”是类别属性。不同铝合金类型下，“杆材强度”与“线径规格”之间的关系可能不同。

将原数据集分为 3 个数据子集，每组数据子集的数据预处理方式类似：

(1) 因为数据量很大，缺失数据量在总样本中的比例很小，故直接删除处理。

(2) 样本集中大部分为重复样本，即多个样本的单线强度和杆材强度相同，重复样本对分析结果没有作用，故删除重复样本。

(3) 在每组数据子集中，以“线径规格”为分组对象，以“杆材强度”为汇总对象，进行分类汇总，计算杆材强度的平均值。以 Excel 为汇总工具，如图 9.6 所示。

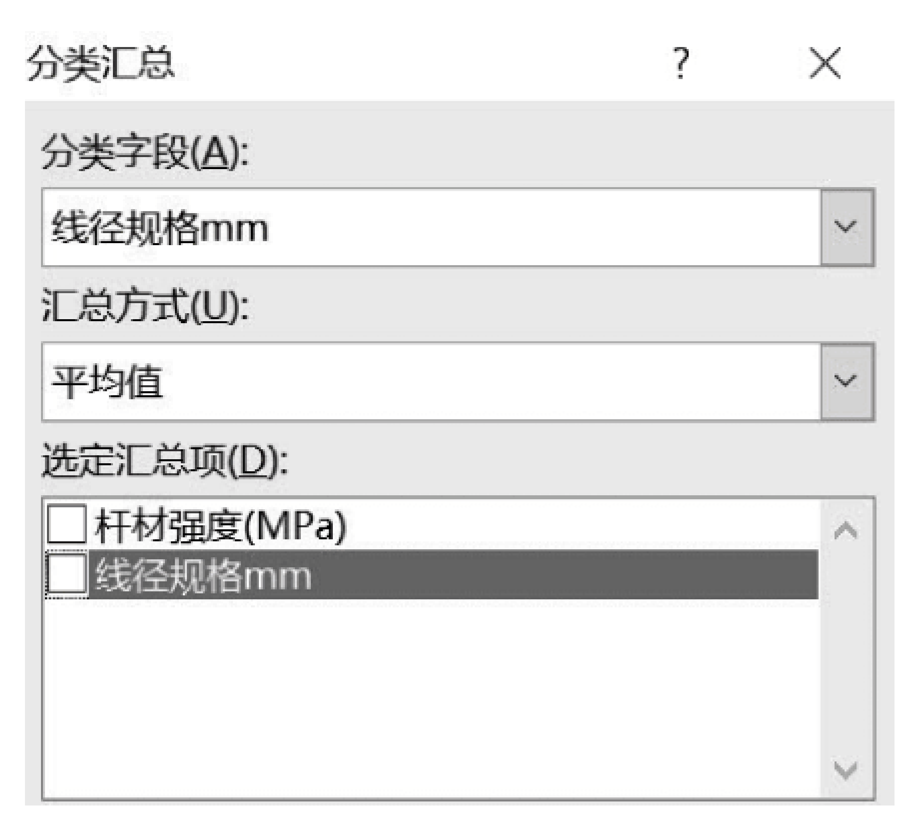


图 9.6　分类汇总

(4) 删除异常值。以“杆材强度”为横轴，以“线径规格”为纵轴画散点图。以 XXX1 型铝合金类型为例，如图 9.7 所示，图中方框中的点为异常点，这里直接删除。

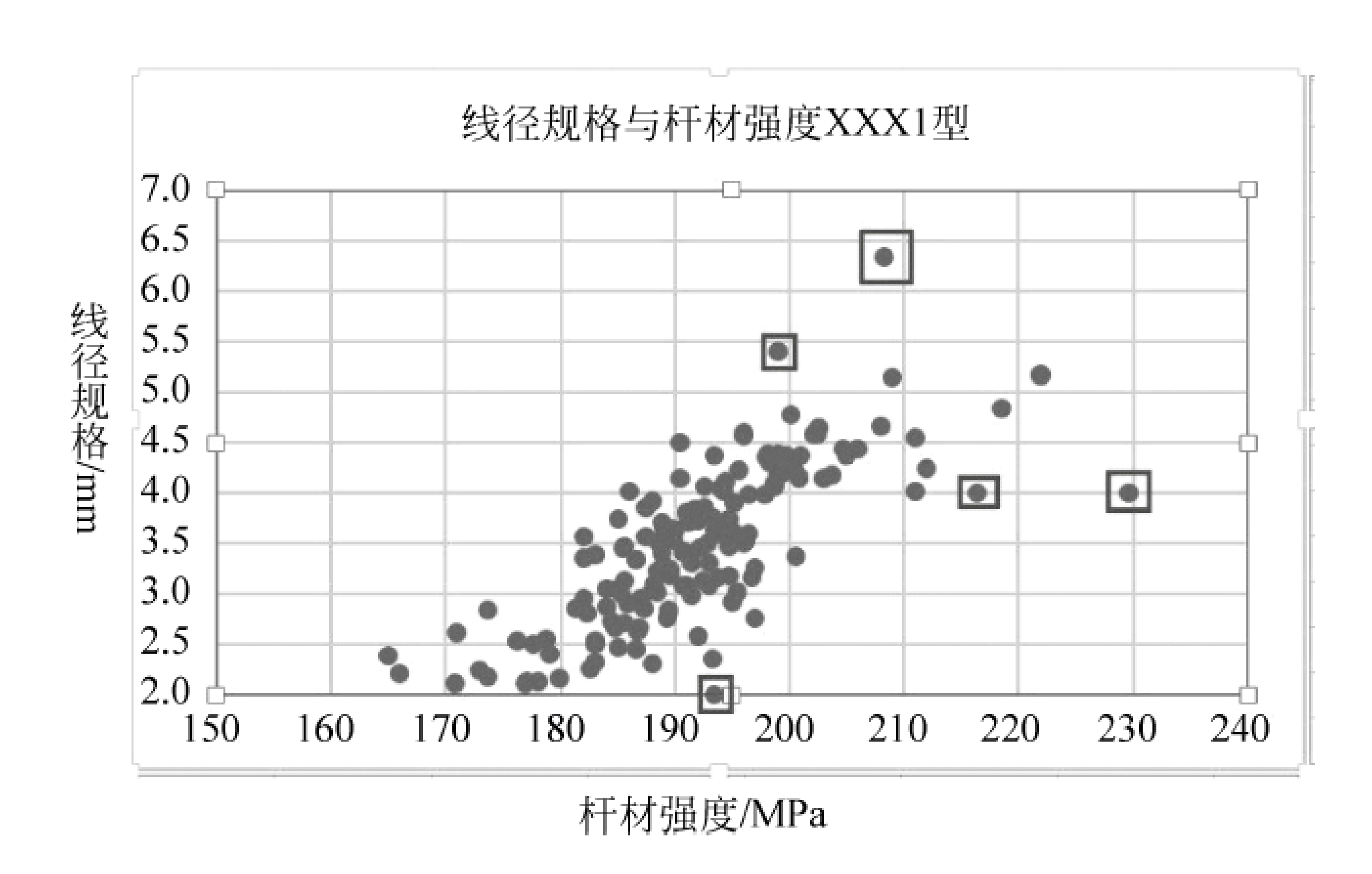


图 9.7 删除异常值

经过上述预处理后，样本数据得到了简化，而且也反映了总体样本的特点。从散点图 9.7 中可以看出，"杆材强度"与"线径规格"大致呈现线性关系，但不是非常明确，从而需要相关系数验证是否存在线性关系以及使用范围区间来表达预测值。

每组数据子集以属性"杆材强度"为因变量 y，以属性"线径规格"为自变量 x 进行回归分析，使用 R 语言分析如下：

```
setwd("D:/Rwork")
options(scipen = 3)
lha1xq <- read.csv('LHA1 - XQ.csv')
with(lha1xq,cor(x,y))
fit1xq <- lm(y~x,data = lha1xq)
summary(fit1xq)
confint(fit1xq,level = 0.90)
```

程序说明如下：

setwd("D:/Rwork")：设置工作表路径为 D 盘的 Rwork 文件夹。

options(scipen=3)：结果表示不用科学记数法。

lha1xq<-read.csv('XXX1-XQ.csv')：读入文件名为 XXX1-XQ 的 CSV 类型文件到 lha1xq 数据集，CSV 文件由 Excel 表转换得来，CSV 类型文件为 R 语言可识别读入文件。

with(lha1xq,cor(x,y))：计算数据集 lha1xq 中 x 与 y 的相关系数。

fit1xq<-lm(y~x,data=lha1xq)：将名为 lha1xq 数据集中的 y 与 x 作线性回归分析，回归函数名为 fit1xq。

summary(fit1xq)：查看回归分析。

confint(fit1xq,level=0.90)：查看 90%的置信区间，表示样本有 90%落在区间范围内。

XXX1 型铝合金的"杆材强度"与"线径规格"相关系数如下：

Cor(x,y)= 0.802 880 5，x 与 y 的相关系数约为 0.80，可以看出具有较强线性关系。但杆材强度与单线线径之间并非简单的线性关系，使用一条线性回归方程并不能完全表达它们之间的关系。为了简化问题，可以使用区间概念近似杆材强度与单线线径之间的非线性关系。

XXX1 型铝合金样本回归分析如图 9.8 所示。

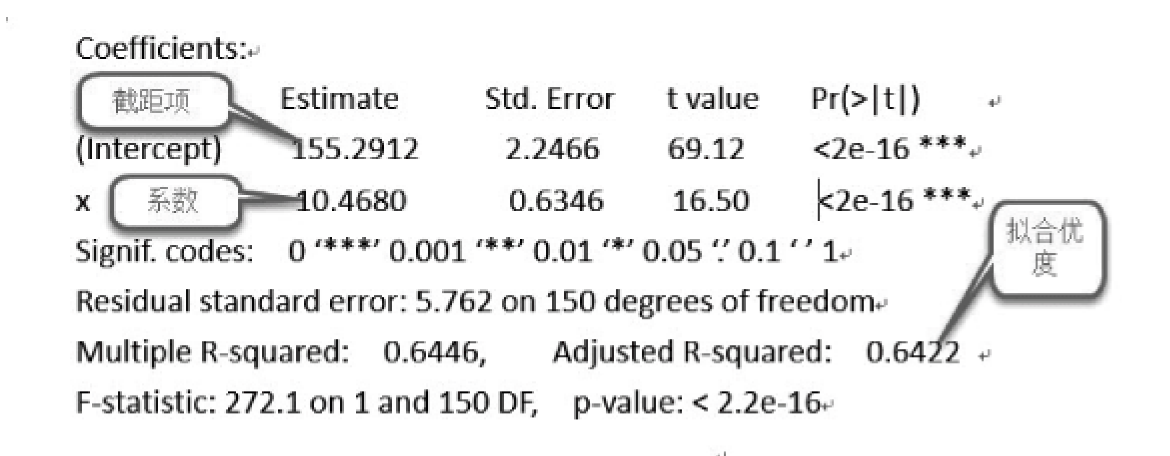


图 9.8　XXX1 型铝合金样本回归分析

从上面的分析结果可见，拟合优度不是很理想，解决的办法是把原有预测为一个具体值的结果改为预测区间的表达方式。这就需要使用上述的 confint 语句。分析结果见表 9.2。预测区间示意图如图 9.9 所示。

表 9.2　XXX1 型铝合金单线线径与杆材强度之间的关系

拟合线信息	回 归 方 程	拟合优度	系数显著性
拟合线	$L: y=155.29+10.47x$	64.22%	***
拟合线上限	$L1: y=159.01+11.52x$		
拟合线下限	$L2: y=151.57+9.42x$		

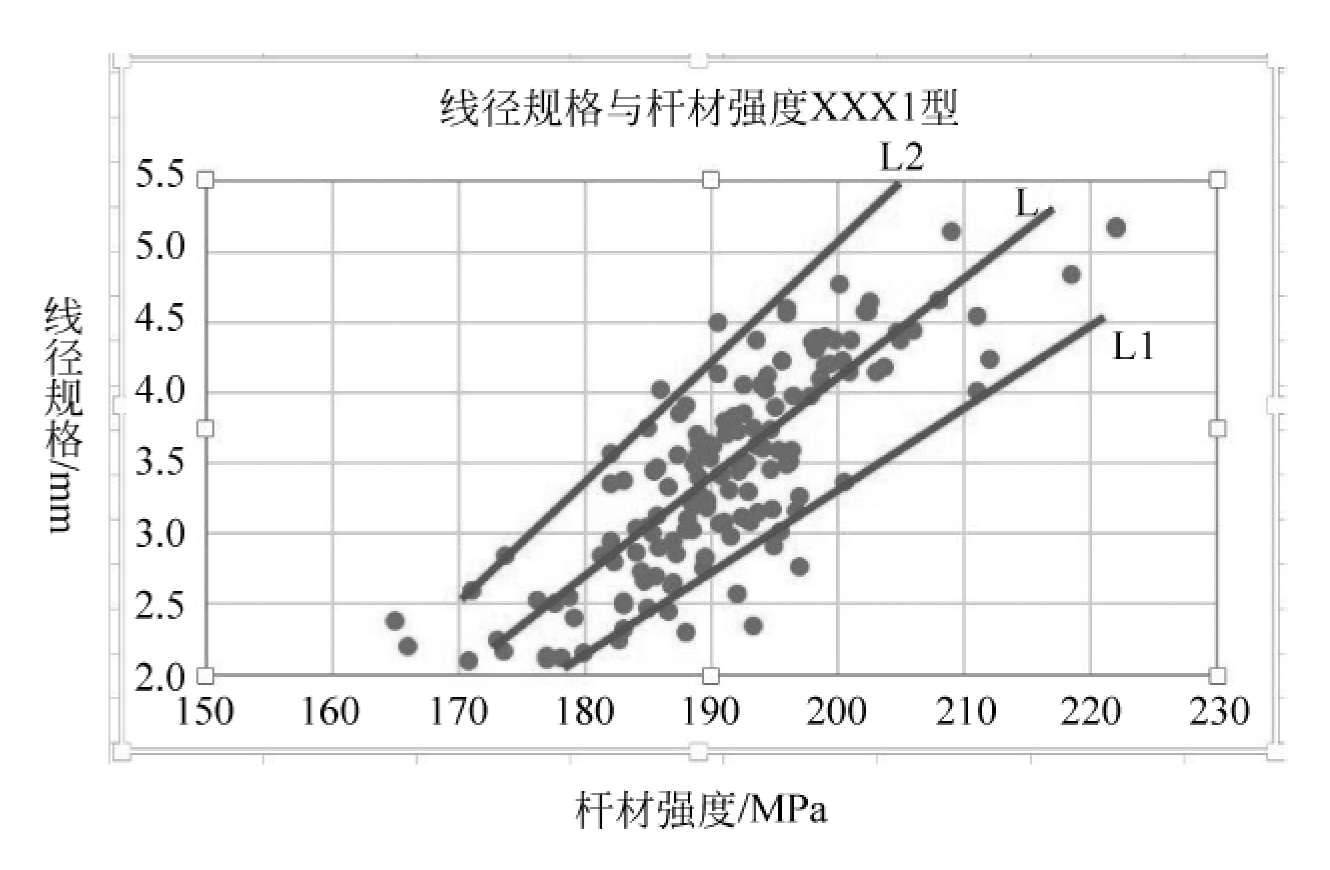


图 9.9　预测区间示意图

根据表 9.2 中的回归分析拟合线下限和上限计算铝合金单线不同线径下的铝合金杆材强度。根据铝合金单线的类型（XXX1）以及线径（2～5mm），将变量 x 代入相应的关系式（表 9.2 中的拟合线下限和上限），计算选取的铝合金杆材强度 y。计算结果见表 9.3。

表 9.3　XXX1 型铝合金单线线径与杆材强度关系表

铝合金线径 XXX1/mm	杆材强度下限/MPa	杆材强度上限/MPa
2.0	170.4083	182.0462
2.5	175.1171	187.8054
3.0	179.8259	193.5645
3.5	184.5348	199.3237

续表

铝合金线径 XXX1/mm	杆材强度下限/MPa	杆材强度上限/MPa
4.0	189.2436	205.0829
4.5	193.9524	210.8421
5.0	198.6613	216.6013

进一步整理得到表 9.4。

表 9.4　XXX1 型铝合金单线线径与杆材强度范围关系

线径/mm	选用杆材强度范围/MPa	流转规定杆材原强度范围/MPa
2.0～2.5	170.41～187.81	170～190
2.5～3.0	175.12～193.56	180～200
3.0～3.5	179.83～199.32	190～210
3.5～4.0	184.53～205.08	190～205
4.0～4.5	189.24～210.84	195～210
4.5～5.0	193.95～216.60	200～215

与原杆材强度选取范围比较，在较细的线径下，杆材强度选取范围有较小程度的减小。在较粗的线径下，杆材选取范围稍有增加。

类似上述思路，还可以分析硬铝的单线线径与所选用杆材强度之间的关系。

通过以上分析，企业可以基于各工序成品质量指标等数据，利用这些数据之间的关联，帮忙耐热导线工厂优化生产流程，降低制造成本，创造最大化获利。

本项目从耐热导线工厂的加工过程数据中发现铝杆、铝线和导线质量指标中的数据异常和规律，从而找到影响加工质量的因素及其内在关系，作为控制整个生产质量的依据，并为设备的预防维修提供决策依据。此外，在一定程度上还可以增强耐热导线工厂利用数据，改进生产流程绩效的意识，并切实利用生产和检测数据不断反思生产工艺中存在的问题，优化生产工艺参数，提高设备的性能，在数字化工厂的基础上建设智能化工厂。

第 10 章

基于逻辑回归模型的高危人员分析

随着社会的开放性、流动性增强，人们的活动范围加大，人口流动性加速，人户分离增多，城市的人口变化加快。特别是随着经济成分、就业方式、组织形式、利益关系和分配方式的多样化，经济社会的活动更加纷繁复杂，大量的“单位人”变成“社会人”，而社会基层组织的社会控制力相对减弱，实有人口管理难度明显加大，城市安全将处于各种机遇和风险的并存期，无论是从维护社会和谐稳定，还是服从服务于经济社会的长远发展出发，加强人口管理工作显得尤为重要。

通过信息化手段加强对违法事件的管理已经成为共识。国外发展较快的地区已经实现了通过数据分析打击犯罪的应用，不少地方已经将数据分析技术引入到实战业务中，形成符合地区特点的数据平台和应用，并取得了实质性的成果。在国内，政府各部门也在奋起直追，通过不断推进无纸化进程，实现数据格式化存储，并不断探索共享和数据应用。

高危人员管理就是充分运用现有的实有人口基础数据和能够反映违法犯罪人员活动规律、行为特点的系统资源，通过建立风险评分模型，实现对人群的比对和分析，排查出具有违法犯罪可疑或可能的高危人员，为派出所民警排查高危人员提供有效的方法和便捷的途径，实现从人到案的打击破案模式，寻找打击破案的增长点，进一步提升人口管理水平和效率，切实做到人口管理更好地为公安实战服务。

对于目前实有人口数量大、社会问题复杂的情况，高危人员探知和对其风险管理手段非常不足。目前，公安部门已经汇集了人口及公安业务多条线的数据，利用这些数据寻找高危人群管控的工作路径成为工作的要点。基于国内外的成功经验，针对现状，整合公安人员轨迹动态数据，进行有效的预处理，形成特征变量后进行归集抽取，并选择逻辑回归模型来构建高危人员评分模型，以此为核心建立高危人员管理系统，通过应用模型逐步满足业务需求。

10.1 高危人员分析需求

高危人员是指在实有人口中有危害国家安全、危害公共安全或其他违法犯罪嫌疑的，须由公安机关进行调查控制，发现、甄别、证实其违法犯罪行为，并依法进行打击处理的人口。该类人员多年来主要靠公安民警和属地政府社会管理人员人工排摸，情况收集，手工上报等方式进行管理。对于高危人员的高危程度，主要依靠办案人员的经验来判定，哪些是高危度较高的，哪些又是关注度较低的。这种方式的操作往往效率和准确性都较低，缺乏科学性和规律性。因此，在公安管理过程中，需要建立一套符合高危人员管理实际的评估体系，综合人员的基本情况、居住情况、就业情况、消费情况、活动场所、社交情况等数据，对人员进行科学合理的高危评估，并对该类人员进行分类和管理。按照人员高危程度采取不同的管控措施，通过细致的分类管理，加强公安对高危人员的管控力度，提高预防等级。

高危人员管理主要解决的问题是，目前高危人员管控缺少手段，预防和排摸都缺少目标性，需要充分挖掘和利用管理方已积累的大量业务数据，以全面分析人员行为和未来发生犯罪之间的关系，将人员高危程度量化，从而科学化地缩小人员管控范围，提高犯罪打击准确度。

嫌疑度的确定上不能简单依靠经验判断，主观臆断更不可取。因此，基于目前人口规模大、重复犯罪率较高、线索往往不足的现状，一套可以将嫌疑人员的嫌疑度直观的量化，并能给予办案人员该量化分值构成及解释的方法，对嫌疑人员范围进行科学的缩小，以此辅助工作开展，就显得极有意义和价值了。

10.2 高危人群相关数据收集与预处理

目前，公安已经实现来沪人员、本市户籍人户分离人员信息在居(村)委中进行采集，得到了大量的数据。如何利用好这些数据，使人口管理工作进一步服务基层实战，从中挖掘出有效、准确、及时和具有指导意义的信息，特别是对于符合公安业务需要的有违法犯罪前科劣迹和需要公安民警重点关注的人员流入情况的提示，以便于民警在工作中对辖区人口成分、结构、层次有更准确、快捷的了解，从而减小民警的工作强度。

所有数据划分为以下几个步骤：原始类型转换、清理、整合、拆分、终止。所有表在数据清理阶段(第一阶段)，判断数据是否重复，如果数据重复，则直接将数据步骤置为终止状态，记录终止原因，所有终止状态数据不参与后面的步骤。

数据清理完成后，根据业务要求将数据整合为人口动态轨迹数据库，并在数据库中根据不同的主题分类抽取数据变量，通过数据导入功能将数据存入数据中心库中。以此完成数据的准备工作，为后续的数据模型的生成建立了基础。

人员轨迹信息同构整合用于将异构的数据源全部同构化到高危人员分析系统数据库中。异构的数据源包括：网吧上网人员的数据、宾(旅)馆、浴场住宿人员数据、违法犯罪人员数据库、吸毒人员数据库、执法办案过程中采集的人员信息、违法犯罪人员手机号码采集系统、看守所释放人员数据、分局查询人员数据、工作对象综合信息系统数据、案事件信息管

理系统数据、服务行业从业人员 IC 卡数据库、实有人口库基础数据等。具体外部数据见表 10.1。

表 10.1　外部数据表

序号	数 据 类 型	数 据 内 容
1	网吧上网人员的数据	人员信息、上网时间、下网时间、网吧名称、网吧地址、所属派出所、经营性质
2	宾(旅)馆、浴场住宿人员数据	人员信息、入住时间、退房时间、场所名称、场所地址、所属派出所、企业名称、营业范围
3	违法犯罪人员数据库	人员信息、案件类型、定罪时间、定罪名称、处理结果、关押时间
4	吸毒人员数据库	人员信息、涉毒类型、是否戒毒、入所时间、出所时间
5	执法办案过程中采集的人员信息	人员信息、采集地点、采集事由、处置结果
6	看守所释放人员数据	全市进看守所人员数据、全市进治安拘留所人员数据、全市刑释解教人员数据、全市刑释强戒人员数据、全市吸毒人员数据、各业务管理确定的工作对象信息
7	违法犯罪人员手机号码采集系统	人员信息、案件类型、定罪时间、手机号码信息
8	分局查询人员数据	人员信息、排查时间、查询事由、人员标签
9	工作对象综合信息系统数据	人员信息、工作对象类型、嫌疑事由、采集时间、处置结果
10	案事件信息管理系统数据	案件类型、案件时间、地点、涉及人员、案件处置结果
11	服务行业从业人员 IC 卡数据库	人员信息、所属单位、管控类别、涉罪情况
12	实有人口库基础数据	姓名、身份证、性别、年龄、户籍地、居住地、职业、学历

导入过程是将数据原样全部以字符串类型导入到数据库，表结构与源文件结构基本一样，增加数据源和导入时间两个字段。导入完成后记录日志，并将源文件从文件缓冲区移动到文件备份区。导入如果失败，则记录错误日志，并向接入监控模块发送警报，将源文件从文件缓冲区移动到文件备份区。全部执行完毕，则开始导入下一个文件，直到文件缓冲区没有文件为止。

进行挖掘的数据必须满足完整性、精确性、一致性等要求，才可以作为数据模型输入的字段值。由于项目的数据来自多个生产系统，不同的系统其数据质量不一，存在数据代码化、关键属性值缺失或无法拆分聚合数据等情况，各数据源的原始数据并未经过加工和处理，需要对数据进行预处理，主要工序包括数据转换、缺损值处理、重复数据处理、噪声数据整理等。

首先，进行数据格式转换，将数据准备库中的原始数据转换成对应的数据类型，并存储在缓冲库中，在数据准备库中根据数据类型分为数值型或日期型，其他数据类型均设置为 NVARCHAR2(2000)。例如，对“宾旅馆入住时间”等日期型字段统一进行日期格式化处理，统一处理为“YYYY-MM-dd hh:mm:ss.ff”形式。缓冲区数据库的字段类型根据数据含义已经设定成了相应的数据类型。

先对缓冲库中的数据表进行扫描，如果有数据，则循环处理每行数据，获取到行数据后，将每个字段的值取出逐个转换，如果全部没有错误，则将该行数据插入到主题库中，并记录操作日志和将原始数据移动到备份表中；如果有错误，则记录错误日志和将原始数据移动到错误表中。缓冲库是数据接入的缓冲区域，原始数据经过数据类型转换后存储在本数据

库中，数据预处理的全过程(数据清理、数据整合、数据拆分、数据转换)将发生在这个数据库中。

对于数据质量较差的字段或进行清洗，或直接抛弃。在源头数据中还发现因采集质量问题而导致的字段值缺失或者是因设计未考虑完整而导致的部分字段值，因在记录中比重较小，所以在数据清洗过程中我们忽略该记录值。其他情况缺失的字段值采用同类属性平均值进行填充，或者可以采用强关联字段的值进行填补。例如，对“违法犯罪人员出所时间”值是空的数据进行统一处理，赋值为“入所时间”加“刑期/拘留时间”；对于模型重要性较小的变量，直接进行剔除操作；对于重要性较高的变量，采取重要性由高到低的方式进行筛选。

对数据噪声的处理，主要采用平滑处理的方式。具体而言，首先对可以选择适宜合并的数据，取该类数据的中间值、边界值、平均值等，对噪声数据进行平滑处理，并对背离度较大的数据予以剔出。在实际处理过程中，对于噪声处理，采用回归方法进行插值处理，对网吧上网次数进行波动分析，其存在连续区间偏移均值较高超过75%的情况，采用回归方法将波动率偏移较大的属性值调低至30%以内。

在重复数据处理方面，由于外部数据源存在相同业务含义数据重叠的情况，如违法犯罪库与工作对象综合信息系统数据库中关于案情描述数据出现记录或属性重复的情况，原因是两种库之间存在继承关系，针对该类情况制订清理规则，明确了重复数据情况下以违法犯罪库为优先，保留该库数据并删除其他相同数据记录。

系统整合的数十个外部数据源中的属性字段值可以组成数以百计甚至更多的可用变量，其中大部分变量与本次数据挖掘工作无关联，如何适当地抽取变量用于模型的创建是非常重要的。首先，在系统初始变量基础上进行相关性分析。例如，分析违法犯罪数据时，发现抓捕时间与案件受理时间相关性较强，因此去除了案件受理时间的初始变量；网吧统计数据中上下网合计次数与上网次数、下网次数相关性较强，因此去除上下网合计次数等变量。因此最后得到清理后的初始变量为67个。然后，根据数据理解和专家讨论，完成了衍射变量的添加。在网吧主题数据中，增加近三个月上网时间在0点之后6点之前的次数、近三个月下网时间在0点之后6点之前的次数等变量；在人口基本主题数据中，增加年龄是否在18～40岁、是否居住于来沪人员倒挂及抓获对象排名最多的居、村等变量；在违法信息主题数据中增加违反犯罪前科次数、涉及案件起数等变量。最后得到的变量个数为20个。

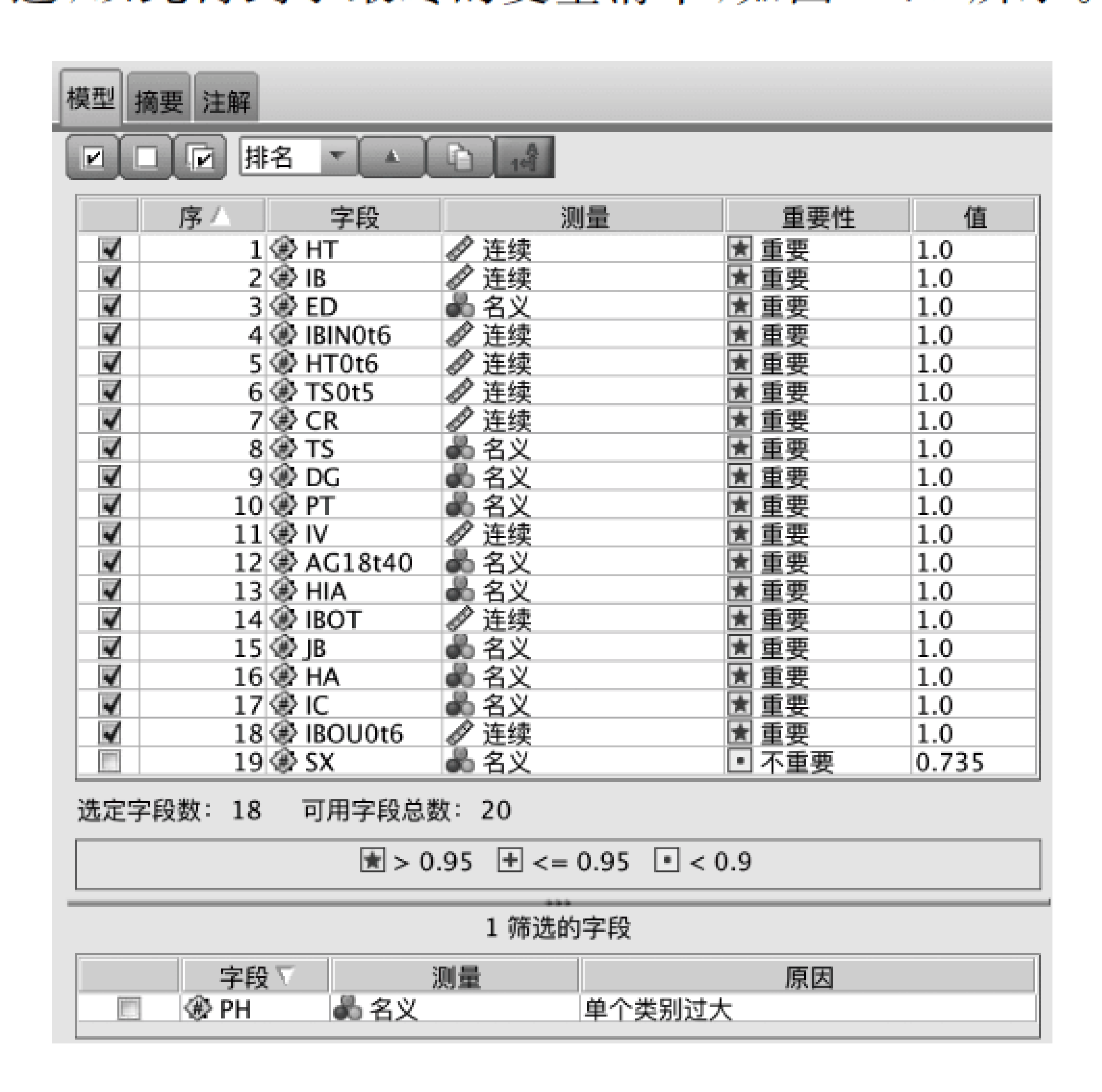
高危人员变量筛选是在初步变量提取的基础上实现模型构建数据准备的重要过程。基于变量筛选中的重要性分析方法进行检验和筛选。

在初始生成的20个变量的基础上进行变量的筛选，其主要思想是对变量进行重要性分析。处理过程主要包括删除强相关性变量以及样本数量较少的对象。然后，通过IBM SPSS Modeler软件的特征选择组件进行二次筛选，在分析方法上使用似然比进行特征重要性分析。

通过对缺失值最大百分比、单个类别中记录最大百分比、最大类别数、最小变异系数、最小标准差等值的设定，划分出重要、边际、不重要等类别的特征变量。

在变量重要性分析中，类别预测变量 p 值(重要性)的基础为“似然比(LR)”。似然比指标可以反映变量真实性，属于同时反映特异度与灵敏度的复合指标。在本检验下，似然比可

以分为犯罪似然比与非犯罪似然比。犯罪似然比为检验结果其高危人员犯罪率和高危人员非犯罪率之比，即检验正确判断高危人员最后成为犯罪人员的可能性与检验错误可能性的比值。其比值越大，则检验结果判断为高危人员成为犯罪人员的概率越大。通过该方法实现了对特征变量的筛选，以此得到了最终的变量清单，如图 10.1 所示。



	序	字段	测量	重要性	值
☑	1	HT	连续	重要	1.0
☑	2	IB	连续	重要	1.0
☑	3	ED	名义	重要	1.0
☑	4	IBIN0t6	连续	重要	1.0
☑	5	HT0t6	连续	重要	1.0
☑	6	TS0t5	连续	重要	1.0
☑	7	CR	连续	重要	1.0
☑	8	TS	名义	重要	1.0
☑	9	DG	名义	重要	1.0
☑	10	PT	名义	重要	1.0
☑	11	IV	连续	重要	1.0
☑	12	AG18t40	名义	重要	1.0
☑	13	HIA	名义	重要	1.0
☑	14	IBOT	连续	重要	1.0
☑	15	JB	名义	重要	1.0
☑	16	HA	名义	重要	1.0
☑	17	IC	名义	重要	1.0
☑	18	IBOU0t6	连续	重要	1.0
☐	19	SX	名义	不重要	0.735

	字段	测量	原因
☐	PH	名义	单个类别过大

图 10.1　筛选后变量数据表

完成了对数据变量的筛选工作后，后续将在人员轨迹数据记录中抽取一个区间段的数据作为样本数据进行模型的训练，创建高危人员评分模型。对抽取的 56 960 条人员轨迹数据进行样本分析、审核和可用性研究，该变量结果数据基本完善，质量较高，可进行后续模型的训练。

10.3　建立模型

逻辑回归是广义线性回归分析模型的一种，在业界已经得到相当广泛的使用。它具有易解释、易使用等优势特点，在公安实际业务中得到充分体现。基于逻辑回归算法的高危人员评分模型通过对人员的基本状况、行为轨迹、前科情况等海量数据全面地予以分析和挖掘，在数据中找寻规律，并以此作为核心来衡量人员的高危程度，为高危人员防控提供了重要基础。

高危分析主要完成分析人员动态轨迹数据与人员犯罪风险情况之间的规律，从而实现对人员高危风险的科学管控。在轨迹数据处理的基础上，使用逻辑回归算法创建高危人员评分模型，实现以高危人员评分模型为计算基础，得到对象人员的高危风险评分，并将该高危评分情况应用至高危人员管理业务中。这里主要探讨基于逻辑回归算法的高危人员评分模型的创建、检验以及高危风险分值转换等功能的设计，并对该应用场景下使用其他算法模型效果进行比较和分析。

完成变量特征筛选后，下面就可以进行模型创建工作了。高危人员评分模型采用 IBM SPSS Modeler 18.0 软件创建模型。高危人员评分模型构建流程如图 10.2 所示。

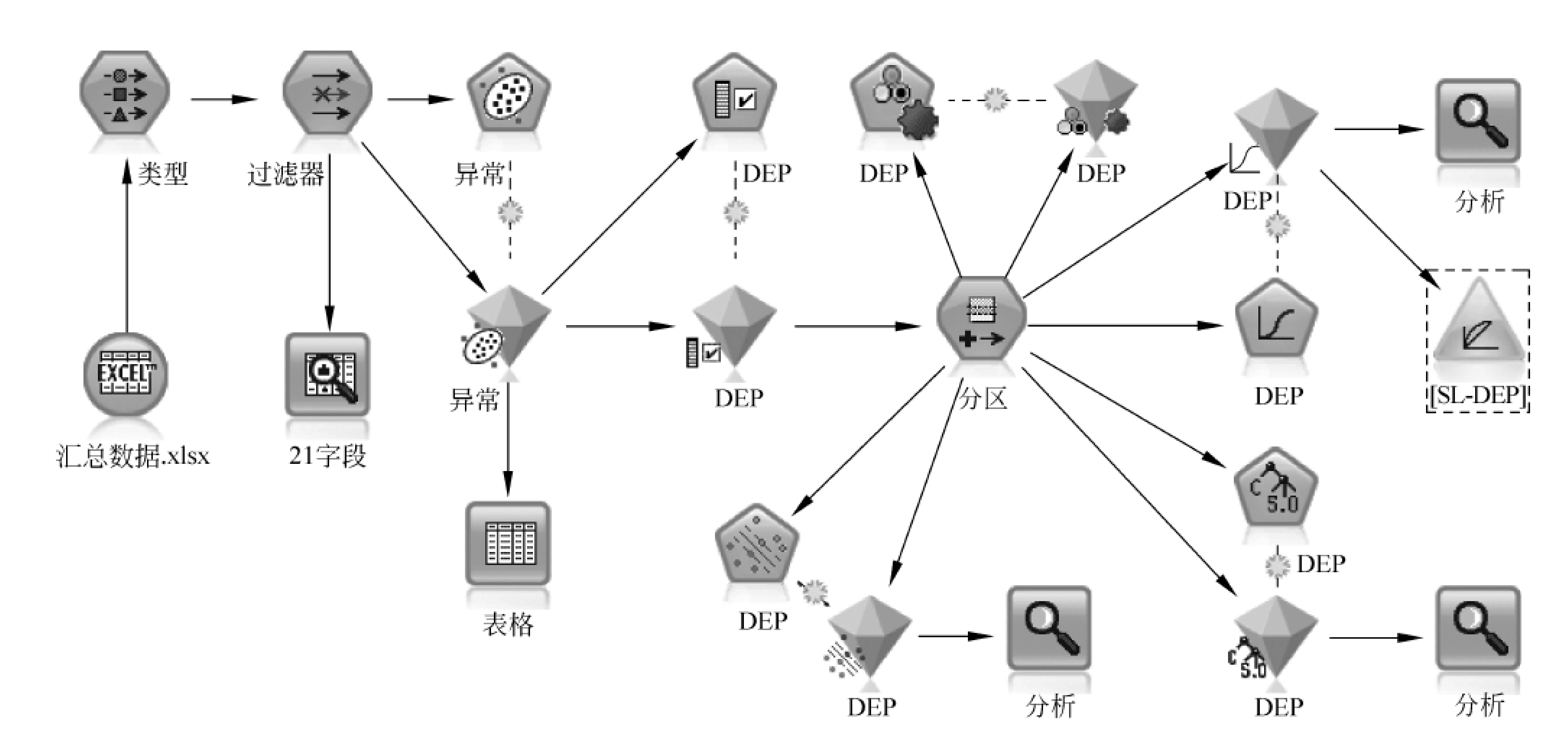


图 10.2 高危人员评分模型构建流程

首先从动态轨迹数据库中的变量表(Label_data)数据对象中获取人员基本信息、网吧宾旅馆信息、违法犯罪信息、关联信息等衍生变量数据,该数据表在之前的 ETL 工作和变量筛选过程中已经清理完毕,该表数据选择的是 2012—2015 年看守所、拘留所入所人员,即这 4 年违法犯罪的人员信息,共计样本数 56960 条。

1. 变量加载

完成数据加载后,使用变量数据筛选工具进行了特征变量的筛选。此外,将 DG(是否存在吸毒史)、JB(有无正当职业)等字段标记为名义类型,汇总数据中的编号、身份证号等字段对分析没有意义,所以将其过滤去除。

2. 异常记录处理

应用异常检测节点对记录中的异常记录进行过滤,分为 3 个对等组,分别得到 41 条、233 条、305 条记录,并将这些记录在流程中舍弃。

3. 数据分区

为保证模型训练的准确度,选择多个数据样本分区方式,原数据中,犯罪人员与未犯罪人员之比为 2∶3,因为数据取样较好,所以犯罪人员信息因为样本少而出现淹没的情况可能性较小。选择数据分区为训练分区占 40%,测试分区占 30%,验证分区占 30%。

4. 模型自动选择

使用“自动分类器”节点并应用分区数据进行模型初选,按照总体精确度进行排序,在“专家”选项卡中,所有模型按照默认参数设置,得到 C5.0、类神经网络、逻辑回归 3 个具有较高总体精确性的模型,如图 10.3 所示。

由于类神经网络在原理解释方面具有较多局限,无法量化说明其模型的依据和原因,所以在本例中不作进一步分析和应用,而逻辑回归具有较强的可操作性,所以优先使用逻辑回归作为标准模型。

高危人员评分中的“是否犯罪”属于二项式分类,因此在回归过程中系统采取二项式方式,并选择向前步进法逐步应用各输入变量,同时选取了专家模式进行参数的调整,对训练

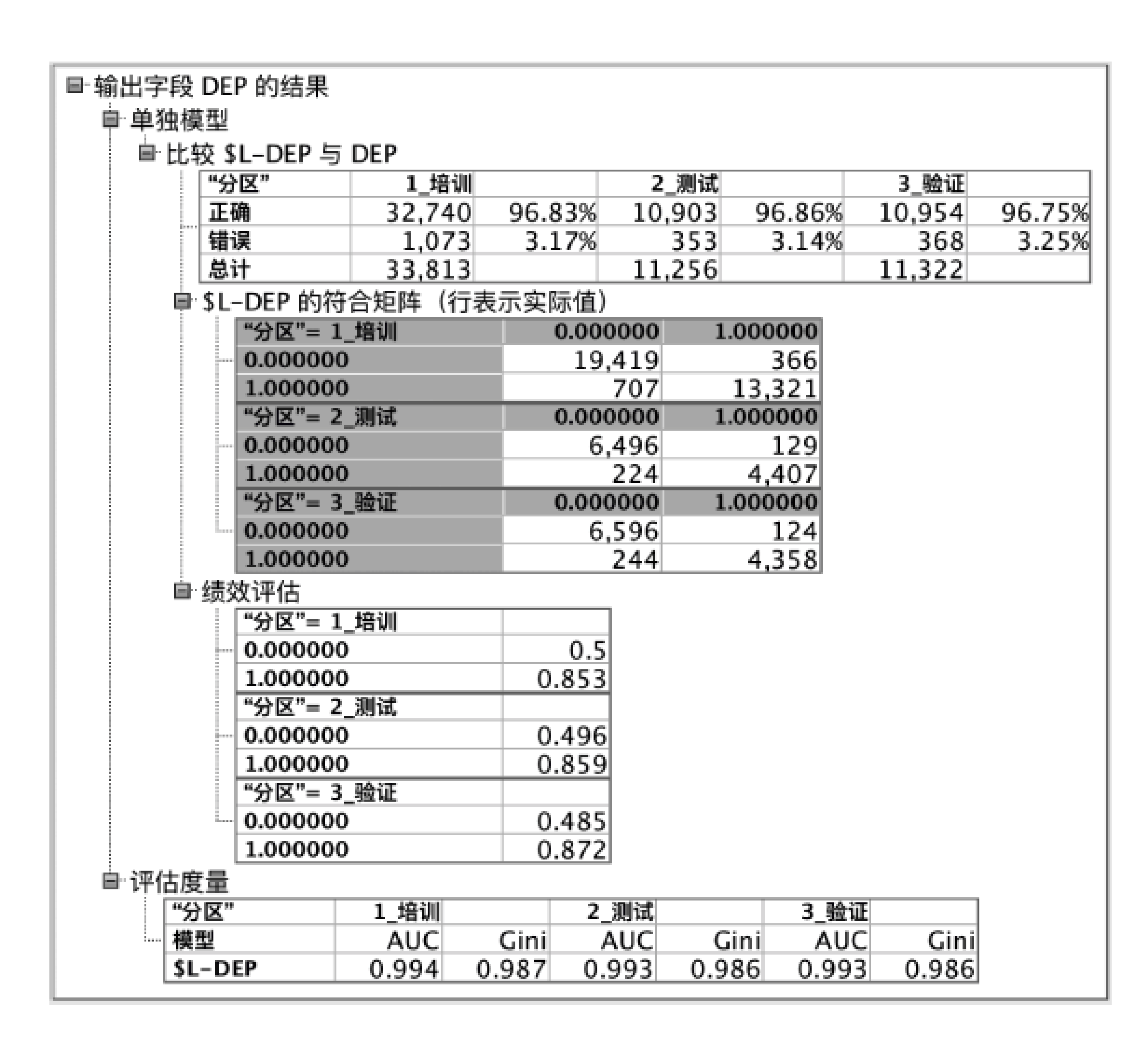
模型　图形　摘要　设置　注解

排序方式(S): 使用　◉升序 ○降序　删除未使用模型　视图: 训练集

图形	模型	构建时间(分钟)	最大利润	最大利润发生比率(%)	增益{前 30%}	总体精确性(%)	使用的字段编号	曲线下方面积
	C5 1	< 1	112,843.3...	40	2.424	98.783	17	0.997
	类神经网络 1	< 1	111,450.0	40	2.423	98.292	18	0.998
	Logistic 回归 1	< 1	107,360.0	39	2.423	96.856	18	0.993

图 10.3　高危人员评分模型构建流程

模式进行进一步的设置，在收敛性条件设置中选择最大迭代数为 25，最大逐步二分法设置为 5。在模型输入中勾选“符合矩阵”“绩效评估”选项，运行后得到模型结果，如图 10.4 所示。



输出字段 DEP 的结果

单独模型

比较 $L-DEP 与 DEP

"分区"	1_培训		2_测试		3_验证	
正确	32,740	96.83%	10,903	96.86%	10,954	96.75%
错误	1,073	3.17%	353	3.14%	368	3.25%
总计	33,813		11,256		11,322	

$L-DEP 的符合矩阵（行表示实际值）

"分区"= 1_培训	0.000000	1.000000
0.000000	19,419	366
1.000000	707	13,321
"分区"= 2_测试	0.000000	1.000000
0.000000	6,496	129
1.000000	224	4,407
"分区"= 3_验证	0.000000	1.000000
0.000000	6,596	124
1.000000	244	4,358

绩效评估

"分区"= 1_培训	
0.000000	0.5
1.000000	0.853
"分区"= 2_测试	
0.000000	0.496
1.000000	0.859
"分区"= 3_验证	
0.000000	0.485
1.000000	0.872

评估度量

"分区"	1_培训		2_测试		3_验证	
模型	AUC	Gini	AUC	Gini	AUC	Gini
$L-DEP	0.994	0.987	0.993	0.986	0.993	0.986

图 10.4　高危人员分析模型

分析模型结果，从准确率来看，具体分析该高危人员评分模型，其中对测试样本的准确性分析为 96.75%，属于犯罪人员的准确性为 95%，不属于犯罪人员的准确性为 98.2%。构建后的高危人员评分模型如图 10.5 所示。

可以从“方程中的变量”表中构建最终的拟合方程式，其中正负号表示的是正相关和负相关，在显著性指标中除了 ED(5)之外，均具有较高的显著性。该拟合方程表示了对象人员存在犯罪可能的几率的自然对数，应用该方程可完成高危人员的高危程度判定，判定如果大于目标阈值，则表明该对象人员可能存在犯罪的可能性，反之，则表明犯罪概率可能性小。

通过拟合方程可以看出，对象人员犯罪与不犯罪之比的自然对数与教育程度、正当职业等变量成反比，与户籍地高危地区、近三月住宿频率、0 至 6 时入住次数、前科数、近三月上网频率、0 至 6 时上网次数等变量成正比，即表明受教育程度越高，正当工作情况存在犯罪可能性越小，而户籍地属于高危地区、前科劣迹、凌晨上网及入住频率出现越高，其犯罪可能性越大。

分类表

实测			预测 DEP 0.0	1.0	正确百分比
步骤 1	DEP	0.0	19419	366	98.2
		1.0	707	13321	95.0
	总体百分比				96.8

模型摘要

步骤	-2 对数似然	考克斯-斯奈尔 R 方	内戈尔科 R 方
1	6319.490[a]	.690	.929

方程中的变量

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)	EXP(B) 的 95% 置信区间 下限	上限
步骤 1[a]	ED			617.295	5	.000			
	ED(1)	1.883	.153	150.522	1	.000	6.574	4.866	8.882
	ED(2)	2.188	.143	234.839	1	.000	8.920	6.742	11.801
	ED(3)	2.632	.151	303.553	1	.000	13.906	10.342	18.699
	ED(4)	1.860	.149	156.757	1	.000	6.427	4.803	8.600
	ED(5)	-.234	.171	1.880	1	.170	.791	.566	1.106
	AG18t40(1)	-.911	.077	138.283	1	.000	.402	.345	.468
	JB(1)	1.171	.071	275.359	1	.000	3.224	2.808	3.702
	HIA(1)	-1.115	.075	218.471	1	.000	.328	.283	.380
	HA(1)	-.836	.080	109.208	1	.000	.433	.371	.507
	HT	.155	.004	1688.316	1	.000	1.168	1.160	1.177
	HT0t6	.159	.005	1013.866	1	.000	1.172	1.161	1.184
	CR	4.504	.228	389.779	1	.000	90.345	57.774	141.278
	IC(1)	.645	.124	27.108	1	.000	1.905	1.495	2.428
	IB	.267	.006	1729.286	1	.000	1.307	1.290	1.323
	IBIN0t6	4.180	.155	731.418	1	.000	65.350	48.272	88.471
	IBOU0t6	.196	.052	14.209	1	.000	1.217	1.099	1.347
	IBOT	-.202	.029	48.238	1	.000	.817	.772	.865
	DG(1)	1.614	.175	85.352	1	.000	5.021	3.565	7.071
	IV	-.363	.051	51.434	1	.000	.696	.630	.768
	PT(1)	1.382	.170	66.476	1	.000	3.984	2.858	5.555
	TS(1)	1.230	.145	72.106	1	.000	3.422	2.576	4.546
	TS0t5	-6.952	.214	1053.603	1	.000	.001	.001	.001
	常量	-9.549	.353	733.549	1	.000	.000		

a. 在步骤 1 输入的变量：ED, AG18t40, JB, HIA, HA, HT, HT0t6, CR, IC, IB, IBIN0t6, IBOU0t6, IBOT, DG, IV, PT, TS, TS0t5。

图 10.5 构建后的高危人员评分模型

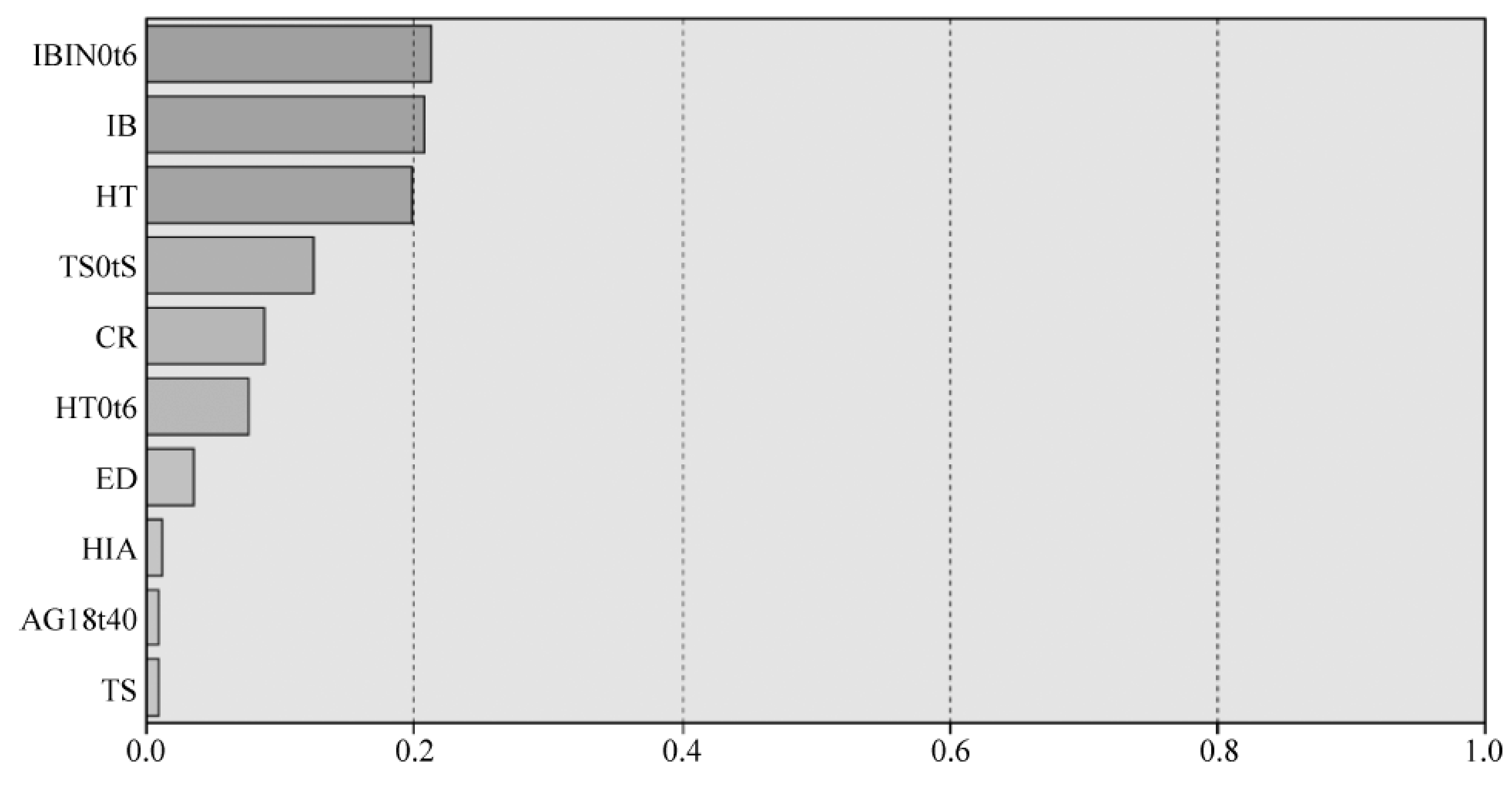


图 10.6 模型预测变量重要性结果

从预测的重要性来看，最近三个月0至6时上网次数、最近三个月上下网次数、最近三个月住宿次数比较关键，其次是最近三个月查询系统的时间段在0至5时之间、犯罪前科次数、最近三个月入住时间在0至6时之间的次数、文化程度，这项的重要性依次减少，其他如户籍地是否为高危地区、年龄在18至40岁之间、是否存在于分局查询系统中这几项影响较弱，剩下的其他因素的重要性极少，几乎不重要。

从结果中可以看到经常在非正常时间上网、住宿的人危险程度较高，需要重点关注此类人群，此外也要关注那些有过前科的较低文化程度的人员，其具有较大重新犯罪的可能性。职业、性别、居住小区是否为来沪人员倒挂和抓获对象排名靠前的居、村等因素并不重要，对结果的影响程度有限。

完成模型训练后，接着需要对模型进行检验评估，利用SPSS Modeler软件增加评估和分析的节点来实现。

5. 模型评估

采用分析节点的模型准确率分析功能，进行模型准确性分析。根据评估功能的结论，对于不同的分区样本，模型的正确率达到了较高的分值，用于验证的样本中，共11 322条数据样本，其中判定正确样本数为10 954条，占总数的96.75%；错误样本数为368条，占总数的3.25%；具体而言，对犯罪人员的判别准确度达到了95%，对非犯罪人员的判断准确度达到了98.2%。即对于验证而言，对某人是否为高危人员采取相关关注措施，判断结果的准确性可以达到95%，对于该类人员可以采取相应的管理手段进行防控，在实际业务中将大大提高公安业务人员的管理手段和管理能力。

在模型效果分析中采用ROC曲线，如图10.7所示，曲线按对象评分高低从低到高排列，纵轴是高危人员的对象累积比例，横轴是非高危人员的对象累积比例，ROC曲线中下面面积表示模型的分辨力。其面积越大，分辨能力越强。从分析图10.7中可以看出，本系统模型的分辨能力较好。

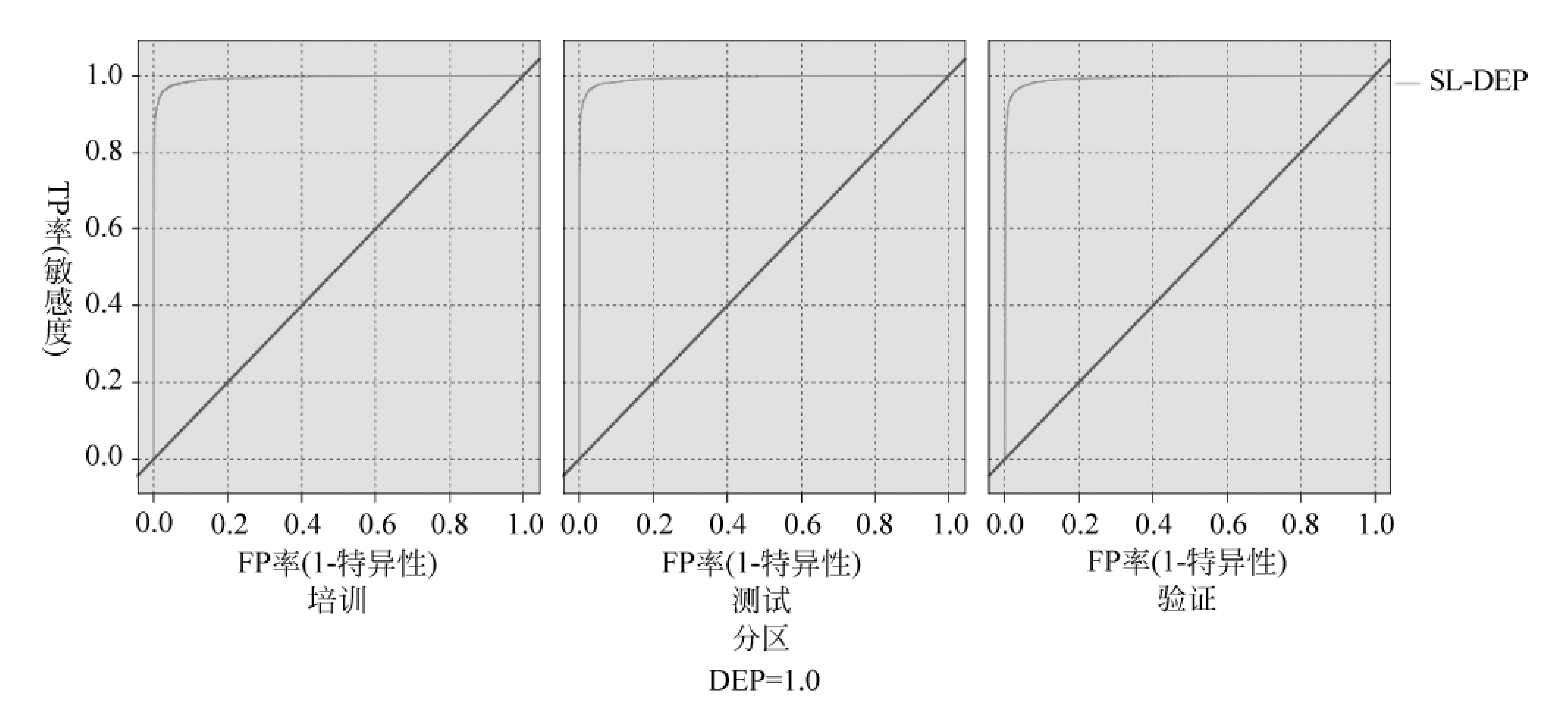


图10.7　高危人员评分模型ROC分析图

在实际应用过程中，由于逻辑回归模型得出的评分为概率系数，在应用中有一定的解释难度，因此需要将概率形式的分数转换为数值型的分数，在系统中，系统利用线形方程对其进行转换，使用计算公式如下：

$$\text{Score} = -\log(\text{odds}) * \text{factor} + \text{offset}$$

利用该公式，系统可以将高危人员分数控制在一定范围内，基准分取为 600，其 odds 是 1∶1，即每当分数增加 15 分，odds 会变成原来的 2 倍，并设定 offset＝600。

通过使用逻辑回归构建的高危人员评分模型可以得出 logit(P)，即对象人员犯罪的可能概率与不发生犯罪的可能概率之比，通过公式 Factor＝15/log(2)将 logit(P)转化为对象人员的高危评分。越高的评分代表对象人员犯罪风险度较高，越低的评分代表其范围风险较小。系统评分模型分值范围在 400～850 分区间的评分模型，600 分为中间分，小于 600 分为高危程度较低对象，可以暂不采用管控措施；大于 600 分为高危程度较高对象，需要采取关注和管控等管理措施，加强监管和防控。

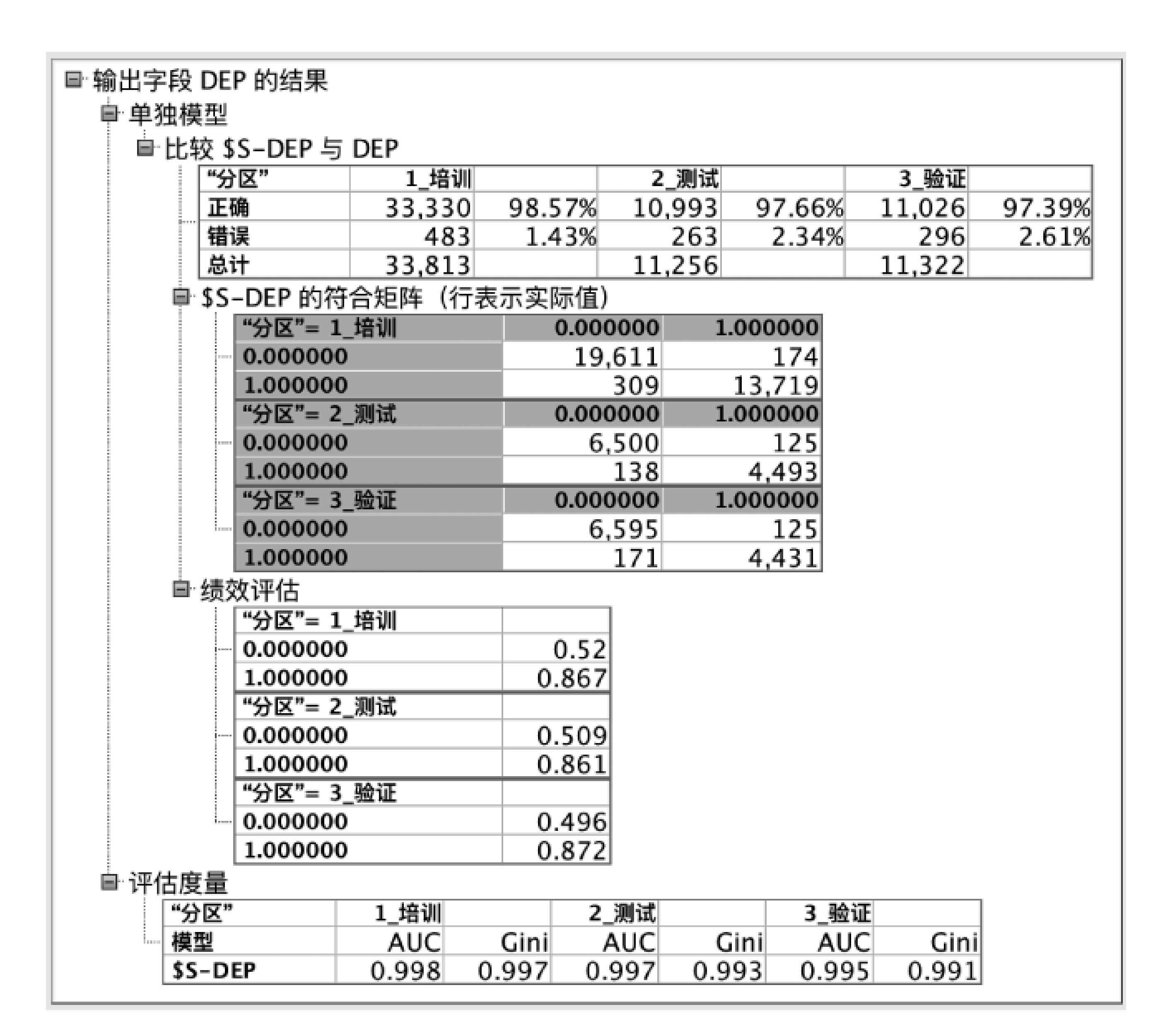
为保证基于逻辑回归的高危人员评分模型的算法有效性，将对其与其他算法的计算结果进行比对和分析。在该分类业务场景中，还可以采用决策树与 SVM，本节对 3 种算法构建的高危人员评分模型进行比较。

采用已构建高危人员评分模型使用的数据样本，在数据理解、数据处理、变量管理等方式均相同的条件下，采用 IBM SPSS Modeler 软件，实现和完成对基于 C5.0 决策树和 SVM 算法的高危人员评分模型。模型构建后，对模型进行检验分析。以下对两种模型的准确性、运行效率以及可解释性进行比较和分析，其模型性能验证结果如图 10.8 所示。

采用 C5.0 决策树和 SVM 算法构建高危人员评分模型后，通过分析模型，得出最后的准确率结果中，逻辑回归算法都是最高的。其中，C5.0 决策树的整体准确率和分类准确率都高于逻辑回归算法；而 SVM 虽然在分类准确率中对于非犯罪人员的判断准确率较高，但是其 AUC 指标相差很小，分别为 0.994 和 0.995，说明两者的准确率等评价指标差别不大，均可满足对高危人员判断的要求，但是，SVM 在可解释性上略差于逻辑回归，所以在这两种方法之间选择，仍然选择逻辑回归模型作为此案例的应用模型。

逻辑回归和 C5.0 决策树在模型理解上都比较有优势，其中逻辑回归对于目标变量与自变量之间过程分析明确，并且对变量值的权值有比较清晰的表达；C5.0 决策树构建过程易于理解，但是随着变量的增多，其节点将会大幅增长，对未来模型的理解会有一定难度。SVM 算法的理论较复杂，对于用户而言，解释和理解具有一定的困难。因此，在公安高危人员管理的业务中，需要业务人员和相关领导快速理解模型，在这一点上逻辑回归算法占有明显的优势。

据此可以综合判定出，逻辑回归算法在预测准确性、运行效率以及可解释性方面都具有一定的优势，因此基于逻辑回归构建的高危人员评分模型是满足高危人员管理业务要求的。



输出字段 DEP 的结果

单独模型

比较 $S-DEP 与 DEP

"分区"	1_培训		2_测试		3_验证	
正确	33,330	98.57%	10,993	97.66%	11,026	97.39%
错误	483	1.43%	263	2.34%	296	2.61%
总计	33,813		11,256		11,322	

$S-DEP 的符合矩阵（行表示实际值）

"分区"= 1_培训	0.000000	1.000000
0.000000	19,611	174
1.000000	309	13,719
"分区"= 2_测试	0.000000	1.000000
0.000000	6,500	125
1.000000	138	4,493
"分区"= 3_验证	0.000000	1.000000
0.000000	6,595	125
1.000000	171	4,431

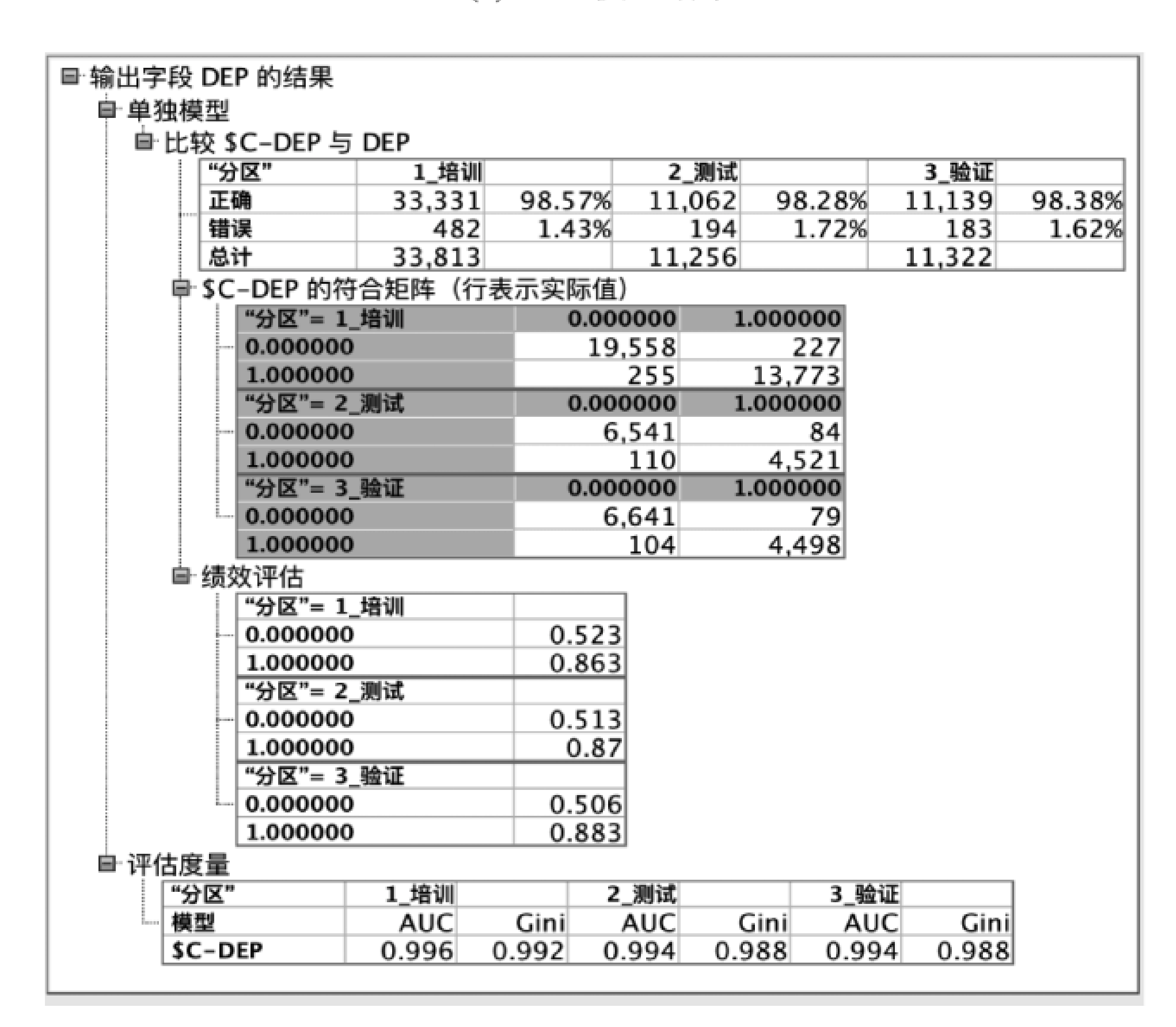
绩效评估

"分区"= 1_培训	
0.000000	0.52
1.000000	0.867
"分区"= 2_测试	
0.000000	0.509
1.000000	0.861
"分区"= 3_验证	
0.000000	0.496
1.000000	0.872

评估度量

"分区"	1_培训		2_测试		3_验证	
模型	AUC	Gini	AUC	Gini	AUC	Gini
$S-DEP	0.998	0.997	0.997	0.993	0.995	0.991

(a) SVM模型结果



输出字段 DEP 的结果

单独模型

比较 $C-DEP 与 DEP

"分区"	1_培训		2_测试		3_验证	
正确	33,331	98.57%	11,062	98.28%	11,139	98.38%
错误	482	1.43%	194	1.72%	183	1.62%
总计	33,813		11,256		11,322	

$C-DEP 的符合矩阵（行表示实际值）

"分区"= 1_培训	0.000000	1.000000
0.000000	19,558	227
1.000000	255	13,773
"分区"= 2_测试	0.000000	1.000000
0.000000	6,541	84
1.000000	110	4,521
"分区"= 3_验证	0.000000	1.000000
0.000000	6,641	79
1.000000	104	4,498

绩效评估

"分区"= 1_培训	
0.000000	0.523
1.000000	0.863
"分区"= 2_测试	
0.000000	0.513
1.000000	0.87
"分区"= 3_验证	
0.000000	0.506
1.000000	0.883

评估度量

"分区"	1_培训		2_测试		3_验证	
模型	AUC	Gini	AUC	Gini	AUC	Gini
$C-DEP	0.996	0.992	0.994	0.988	0.994	0.988

(b) C5.0模型结果

图 10.8　模型性能验证结果

第 11 章

卷积神经网络在音频质量评价领域的应用

深度学习是机器学习的一个重要分支，是目前数据科学领域中比较热门的研究方向，其起源于神经网络，随着近几年计算能力的提升和大数据的快速应用逐渐发展起来，人工智能领域的很多应用都采用了深度学习相关的理论和技术，特别是在自然语言处理、计算机视觉、图像识别、场景分类等方面成果显著。人工智能的应用获得大众普遍关注后，深度学习相关技术已成为数据分析人员的必修目标之一。本章主要从深度学习的理论基础、发展历程、常用算法等几个方面对其进行介绍，并结合案例说明其在音频质量评价方面的应用。

11.1 深度学习基础

本节主要阐述深度学习相关理论基础，对比人工神经网络的特点说明深度学习的基础概念和训练过程，通过介绍深度学习的发展历程可以从纵向角度来把握深度学习的发展脉络，可对未来发展趋势有更深刻的理解。另外，本节还介绍了目前主流的技术框架及其特点，可以在实践中按业务需求选择合适的框架进行应用。

基于反向传播算法(Back Propagation，BP)的传统人工神经网络是一种浅层学习模型，由于运算能力的限制，往往只有输入层、隐含层、输出层，容易产生过拟合，泛化能力较差。深度学习的基本思想是通过使用多个层，某一层作为下一层的输入，来实现对输入信息的分级表达，这参考了人类的分层处理系统，可以让机器自动地学习有用的特征，采用多层神经网络的结构来抽象特征，从而发现更多的数据分布特点。

深度学习的目标是模拟人类大脑进行学习，通过多个层对特征进行学习，特征表示的粒度要具有一定的结构性，不仅在横向的维度中具有关联，而且要在纵向抽象时具有意义，从特征的稀疏编码逐渐迭代抽象，复杂度和抽象度逐层递增，而抽象的层次越高，其类别越少，也就更易于区分。可以说，深度学习就是一种非监督式特征学习的过程。

深度学习的训练过程是按照分层训练的机制，自底向上进行非监督特征学习，获得各层的参数，也可以认为是对相应特征进行学习的过程。当然，其偏差也会逐层传递。在达到最顶层之后对比结果标签，对误差自顶向下逐层传输，进行有监督学习，对各层中的参数进行微调，通过多次迭代调整，使整个网络的参数具有较好的区分效果。

11.1.1 深度学习的发展过程

人工神经网络经过最近几十年的发展，从1943年心理学家McCulloch和数学家Pitts参考生物神经元的结构发明了神经元模型之后，从单层神经网络到两层网络，再到多层神经网络，随着层数的增加和激活函数的不断演变发展，其非线性拟合能力不断加强。随着计算的运算能力和数据量几何级的增长，以及更多训练模式的引入，神经网络在人工智能领域发挥着越来越大的作用。

日本的Fukushima于1980年第一次提出基于感受野的模型。1998年，由Lecun等人提出的LeNet-5卷积神经网络模型用于对手写字母进行文字识别，它是基于梯度的反向传播算法对模型进行训练，将感受野理论应用于神经网络中。

2006年，多伦多大学的G. E. Hinton等提出深度学习的概念。深度学习是一种多层级的深层次网络结构的机器学习方法，主要是为了解决传统的神经网络很容易收敛到局部最小值这一问题，Hinton提出使用无监督预训练的方法优化网络权值的初值，再进行反向参数调整的方法来优化网络性能。

2010年，深度学习项目首次获得来自美国国防部门DARPA计划的资助，参与方有美国NEC研究院、纽约大学和斯坦福大学。自2011年起，谷歌和微软研究院的语音识别方向研究专家先后采用深度神经网络技术将语音识别的错误率降低20%～30%，这是长期以来语音识别研究领域取得的重大突破。2012年，深度神经网络在图像识别应用方面也获得重大进展，在ImageNet评测问题中将原来的错误率降低了9%。2012年6月，Andrew NG等对机器进行大量训练以后，使其学会自动识别猫的图像。

2014年，Ian Goodfellow将生成对抗网络(Generative Adversarial Networks，GAN)引入深度学习领域。2016年，GAN热潮席卷AI领域顶级会议，从ICLR到NIPS，大量高质量论文被发表和探讨。2016年3月，Google公司的AlphaGo战胜韩国顶尖围棋棋手李世石，2017年1月4日，又以Master为账号，在未公开身份的情况下，通过网上比赛战胜了中韩日台的顶尖围棋手60多人，而AlphaGo采用的神经网络技术中就包括了卷积神经网络和生成对抗网络。

卷积神经网络已经成为当前深度学习领域的热点，特别是在图像识别和模式分类方面，其优势是共享权值的网络结构、局部感知(也称为稀疏连接)，降低神经网络的运算复杂度，因为减少了权值的数量，并可以直接将图像作为输入进行特征提取，避免了对图像的预处理和显式的特征提取，可以进行同步学习。与之相关的是循环神经网络(RNN)、长短期记忆网络(Long Short Term Memory networks，LSTM)等。

11.1.2 深度学习常用技术框架

目前，深度学习领域中的主要实现框架有Torch、TensorFlow、Theano、Caffe、Keras、MxNet、Deeplearning4j等，下面详细介绍各框架的特点。

1. Torch

Torch 是用 Lua 语言编写的带 API 的深度学习计算框架，支持机器学习算法，其核心是以图层的方式定义网络，优点是包括了大量模块化的组件，可以快速进行组合，并且具有较多训练好的模型，可以直接应用。此外，Torch 支持 GPU 加速，模型运算性能较强。

Torch 虽然功能强大，但其模型需要 LuaJIT 的支持，对开发者学习和应用集成都具有一定的障碍，文档方面的支持较弱，对商业支持较少，大部分时间需要自己编写训练代码。目前最新的 Torch 是由 Facebook 在 2017 年 1 月正式开放了 Python 语言的 API 支持，即 PyTorch，支持动态可变的输入和输出，有助于 RNN 等方面的应用。

2. TensorFlow

TensorFlow 是用一个 Python API 编写的，通过 C/C++引擎加速，由谷歌公司开发并开源，影响力较大且社群用户数量多，对应的教程、资源、社区贡献也较多，出现问题后更易查找解决方案。它不止用于深度学习，还支持强化学习和其他算法的工具，与 NumPy 等库组合使用可以实现强大的数据分析能力，支持数据的并行运行和模型的并行运行，在数据展现方面，可以使用 TensorBoard 来对训练过程和结果按 Web 方式进行可视化，只要在训练过程中将各项参数值和结果记录于文件中即可。

TensorFlow 的主要缺点是在性能上较 Torch 等框架差一些，也比 Torch 笨重一些，较难理解，其动态类型在大型项目中容易出错，不利于工具化，且不提供商业支持。

3. Theano

Theano 是早期的深度学习框架，用 Python 编写，其应用级别较低，深度学习领域的许多学术研究者较多地使用它。Theano 可与其他学习库配合使用，非常适合数据探索和研究活动。其在大型模型上的编译时间较长，启动时间较长，只支持单个 GPU，实际项目应用中局限性较多。

现在像 Keras 这样比较流行的开源深度学习库，都是在 Theano API 的基础上进行开发的，目前对 Theano 感兴趣的开发者越来越少，与之相关的库有的已经停止更新了，所以目前并不适合应用开发人员使用。

4. Caffe

Caffe 是较早的一个应用较广的工业级深度学习工具，将 Matlab 实现的快速卷积网络移植到了 C 和 C++平台上。它不适用于文本、声音或时间序列数据等其他类型的深度学习应用，在 RNN 方面建模能力较差。Caffe 选择了 Python 作为其 API，但是模型定义需要使用 protobuf 实现，如果要支持 GPU 运算，需要自己用 C++/CUDA 来实现，用于像 GoogleNet 或 ResNet 这样的大型网络时比较烦琐。Caffe 代码更新趋慢，可能未来会停止更新。

5. Keras

Keras 是由谷歌软件工程师 Francois Chollet 开发的，是一个基于 Theano 和 TensorFlow 的深度学习库，具有较直观的 API。这可能是目前最好的 Python API，未来可能会成为 TensorFlow 默认的 Python API，其更新速度较快，相应的资源也较多，受到广大开发者追捧。

6. MxNet

MxNet 是一个提供多种 API 的机器学习框架，主要面向 R、Python 和 Julia 等语言，由华盛顿大学的 Pedro Domingos 及其研究团队管理维护，具有详尽的文档，容易被初学者理解和掌握。它是一个快速灵活的深度学习库，目前已被亚马逊云服务采用。

7. Deeplearning4j

Deeplearning4j 是用 Java 编写的，所以可用性较好，对开发人员来说，学习曲线较低，在现有的 Java 系统中集成使用更加便利。通过 Hadoop、Spark、Hive、Lucene 等这类的开源系统来扩展可实现无缝集成，具有良好的生态环境支持。Deeplearning4j 中提供了强大的科学计算库 ND4J，可以分布式运行于 CPU 或 GPU 上，并可通过 Java 或 Scala 进行 API 对接。Deeplenrning4j 与 Caffe 类似，也可以快速应用 CNN、RNN 等模型进行图像分类，支持任意芯片数的 GPU 并行运行，并且提供在多个并行 GPU 集群上运行。

Deeplearning4j 提供了实时的可视化界面，可以在模型训练过程中查看网络状态和进展情况。当然，使用实时查看功能时将影响模型训练的性能。

11.1.3　常用的深度学习算法

本节将详细介绍几种常见的深度学习算法，包括卷积神经网络、循环神经网络、生成对抗网络(Generative Adversarial Network，GAN)，这几种算法为深度学习的基础算法，在各种深度学习相关系统中均有不同程度的应用。除此之外，目前比较前沿的深度学习算法还有自动机器学习(Auto Machine Learning，AutoML)，其中代表项目为 AutoML，可以帮助我们尝试各种不同的算法并选择最佳算法，然后进行超参数调优，并可以对模型结果进行评估。

1. 卷积神经网络

卷积神经网络是一种比较常见的深度学习算法，是一种监督式学习的深层神经网络，由于它稀疏的网络结构，在层的数量、分布、每一层卷积核的数量都会有差异，结构的好坏决定了模型运算的效率和预测的精确度。理解不同结构层次的作用和原理有助于设计符合实际的深层网络结构。

卷积层和子采样层是特征提取功能的核心模块。卷积神经网络通常采用梯度下降的方法，应用最小化损失函数对网络中各节点的权重参数逐层调节，通过反方向递推，不断地调整参数，使得损失函数的结果逐渐变小，从而提升整个网络的特征描绘能力，使网络的精确度和准确率不断提高。

卷积神经网络前面几层由卷积层和子采样层交替组成，在保持特征不变的情况下减少维度空间和计算时间，更高层次是全连接层，其输入是由卷积层和子采样层提取到的特征，最后一层是输出层，可以是一个分类器，采用逻辑回归、Softmax 回归、支持向量机等进行模式分类，也可以直接输出某一结果数值。经典的 LeNet-5 卷积神经网络结构图如图 11.1 所示，其中包括以下几个主要的层次结构。

1）卷积层

通过卷积层(Convolutional Layer)的运算，可以将输入信号在某一特征上加强，从而实现特征的提取，也可以排除干扰因素，从而降低特征的噪声。

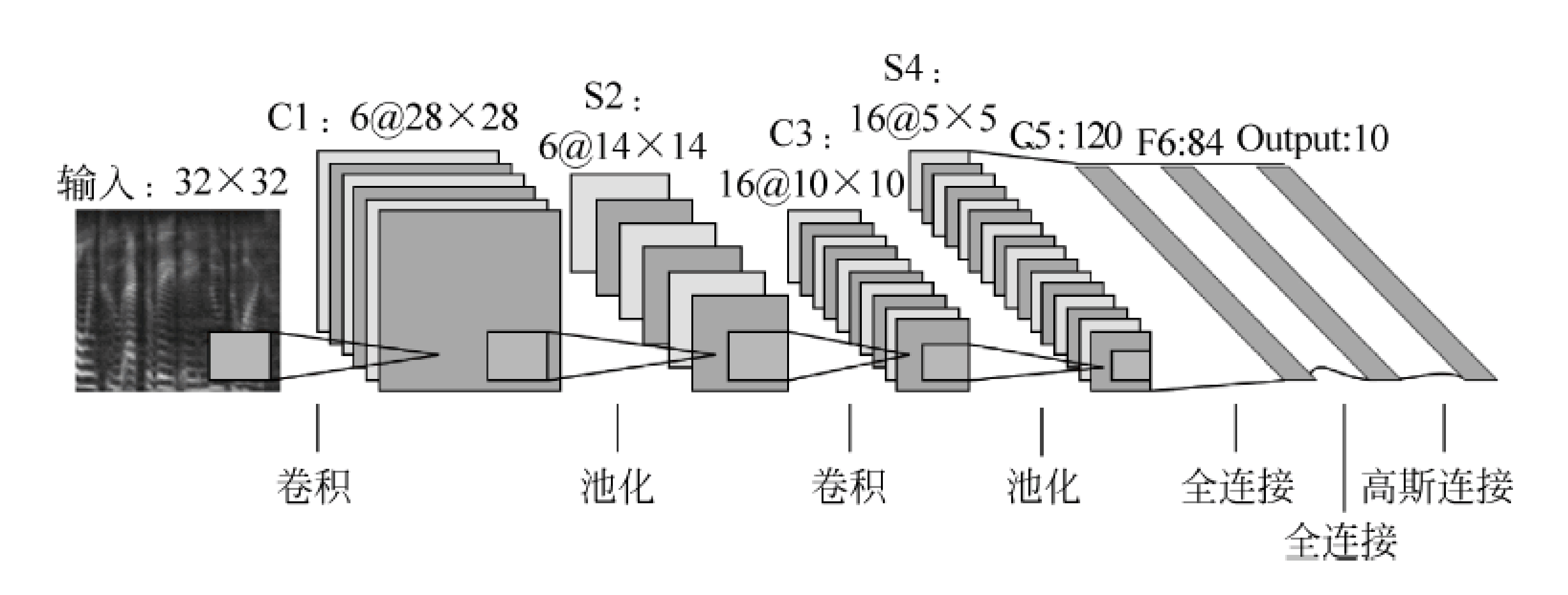


图 11.1 经典的 LeNet-5 卷积神经网络结构图

2）线性整流层

引入 ReLU 层(Rectified Linear Units Layer)的主要目标是解决线性函数表达能力不够的问题。线性整流层作为神经网络的激活函数(Activation function)可以在不改变卷积层的情况下增强整个网络的非线性特性，在不改变模型的泛化能力的同时数倍地提升训练速度。线性整流层的函数有以下几种形式：

$$f(x)=\max(0,x)$$
$$f(x)=\tanh(x)$$
$$f(x)=|\tanh(x)|$$
$$f(x)=(1+e^{-x})^{-1}$$

其中，$f(x)=(1+e^{-x})^{-1}$，是 Sigmoid 函数，是传统的神经网络激活函数，将实数压缩在 0～1，这样就可以将其用于分类的操作。但在实际梯度下降中，容易出现梯度消失，导致终止梯度传递，所以目前主要使用 ReLU 函数 $f(x)=\tanh(x)$作为激活函数，优点是收敛快，并且计算成本低，原因是它模仿了生物学的原理，研究表明生物神经元的信息编码是比较分散和稀疏的，能有效地进行梯度下降和反向传播，可以避免梯度消失的问题，同时，活跃度的分散性使得网络的运算成本较低。

3）池化层

池化层(Pooling Layer)是一种向下采样的形式，在神经网络中也称为子采样层(Sub-sampling Layer)，一般使用最大池化(Max Pooling)将特征区域中的最大值作为新的抽象区域的值，减少数据的空间大小，所以参数的数量和运算量也会减少，减少了全连接的数量和复杂度。这一理论的基础是特征的相对位置比具体的实际数值或位置更加重要，所以是否应用池化层需要依照实际需要进行分析，否则会影响模型的准确度。

4）全连接层

卷积层得到的每张特征图表示的是输入信号的一种特征，而它的层数越高，表示这一特征越抽象，为了综合低层的各个卷积层特征，就加上全连接层(Full Connect Layer)将这些特征结合到一起，然后用 Softmax 等进行分类或逻辑回归分析。

5）输出层

输出层(Output Layer)的一项任务是进行反向传播，依次向后进行梯度传递，计算相应的损失函数，并重新更新权重值。在训练过程中可以采用 Dropout 避免训练过程产生过拟合。输入层的结构与传统神经网络结构相同，是基于上一全连接层的结果进行类别判别。在实际应用中具有多少个标签分类，在输出结果时就设置多少个输出。

2. 循环神经网络

循环神经网络分为时间循环神经网络和结构循环神经网络，通常指的是前一种，之所以是"循环"，是因为其中隐藏层节点的输出不仅取决于当前输入值，还与上一次的输入相关，即节点的输出可以指向自身，进行循环递归运算，在处理时间序列相关的场景时效果明显，因为每个观察样本都与之前的样本关系密切，所以其在分析语音、视频、天气预报、股票走势预测等方面具有突出优势。

RNN存在的问题是在处理长时间关联关系时，要记住所有的历史样本参数，复杂度增加，容易导致权重参数出现梯度消失或梯度爆炸。为避免此类问题，一般采用长短时记忆(Long Short Term Memory，LSTM)网络来处理，原理是其神经元的结构与传统神经元不同，称为记忆细胞(Cell State)，其包括了输入门(input gate)、遗忘门(forget gate)、输出门(output gage)，在循环过程中，元胞状态接受输入数据的影响，在遗忘门里更新记忆状态，并将其通过输出门进行输出，其关键在于应用遗忘门将重要的因素进行记录，减少了记忆的元素数量，使得在模型训练时具有较强的梯度收敛性。

3. 生成对抗网络

传统的深度学习通常需要大量的样本进行训练，如果是进行监督式学习的方法，需要人工进行样本标记，费时费力。为了解决这一问题，可以通过自动编码器(Auto Encoder)、受限玻尔兹曼机(Restricted Boltzmann Machine，RBM)、深度置信网络(Deep Belief Network，DBN)等方法实现非监督式学习样本特征，另外一种方法是使用生成对抗网络，它解决的问题是从现有样本中学习并创建出新的样本，按照人类对事物的学习过程，逐渐总结规律，而并非大数据量地训练，所以在新的任务处理中，只需要少量的标记样本就可以训练出高效的分类器。

GAN网络中需要两个神经网络：一个是生成网络G，另外一个是区分网络D，前者的主要任务是生成新的样本，后者的主要任务是对样本进行区分，首先训练区分网络D，从而提高模型的真假分辨能力，然后训练生成网络G，提高其欺骗能力，生成接近于真实的训练样本。两种网络之间形成对抗关系，都极力优化自己的性能，直到达到一种动态平衡状态，使得区分网络难以区分(准确率为50%)。

在两种网络训练过程中需要注意，在某些时候，G网络容易简单生成与训练集中样本相差不大的新样本，导致D网络无法区分，实际上，新样本中种类的数量不多，为了避免此类过拟合，可以在D网中计算样本间的相似度，并作为特征传入下一层中，这样就可以识别出假的样本，从而进行惩罚，促使G网络生成多种新样本。

另外，如果D网络过于强势，可能会导致G网络中参数梯度较大，无法有效收敛，可以在训练过程中调低训练样本的概率目标，这种方法也称为单边标签平滑。

11.2　音频质量评价

随着移动互联网的兴起和自媒体音频服务的流行，人们对媒体的质量要求越来越高，并且要求对音频质量评价快速和稳定。本节将从音频质量的评价标准和影响因素出发，应用卷积神经网络模型对音频质量进行评价，将低层音频特征、梅尔倒频谱系数、语谱图等特征

信号作为输入参数传入卷积神经网络，应用感知模型对信号进行卷积、池化等综合计算并映射为音频质量分类结果。

为了验证卷积神经网络的应用效果，使用 Deeplearning4j 作为深度学习计算框架。另外，音频特征提取采用 librosa 来对音频特征进行提取，并将提取到的结果进一步处理为图片格式，作为输入传递到卷积神经网络中进行实验。

实验需要安装的基本软件为：Java 为 64 位，且版本为 1.7 以上；Apache Maven，主要用于依赖包的自动管理；Intelli IDEA 或者 Eclipse，推荐使用前者；Git，用于代码管理；Python 环境，版本为 2.7，需要安装 pip 支持，用于安装 librosa 库。除此之外，还要安装 numpy、matplotlib、scipy、sklearn、PIL；ffmpeg，Mac OSX 可使用 homebrew 安装，Windows 用户需要单独安装。

11.2.1 音频样本及特征预处理

从国内某自媒体平台中通过人工听评的方式，随机选择 200 个音频作为音频集，并按高质量、一般质量、低质量分为 3 类，然后将其按 4∶1 的比例分为两部分，20%的音频作为验证集音频，80%的音频作为训练音频库。

对音频特征进行预处理包括特征提取和特征选择。与音频质量相关的特征主要有音频低层特征、MFCC 特征、心理声学特征。音频特征进行提取并将其应用到模型中进行验证，并确认选取特征。

音频质量客观评价分析过程中，需要确认选择哪些指标进行模式学习，选择指标的过程就是特征提取的过程，包括音频低层特征、MFCC 特征、心理声学特征，对提取的 MFCC、Spectrogram 等特征存储为图片格式。

1. 低层特征

音频低层特征是可以直接通过时域波形或频域信号中对每一音频帧进行加窗运算获得，这些特征已经广泛应用于音频处理应用中，如语音识别、音乐分类，甚至应用于通过分析设备运行的声音来识别其故障种类，在音频质量评价中也将引用特征进行基本的音频质量分析。

使用 librosa 对音频、低层特征 RMSE 进行提取，具体示例代码如下。如果要提取其他特征，可以选择对应 librosa.feature 库中的特征，如将 librosa.feature.rmse 中的 rmse 替换为 spectral_centroid、spectral_bandwidth、zero_crossing_rate 等，来提取频谱质心、频谱带宽、过零率等特征。

```
import librosa
import librosa.display
y,sr = librosa.load('./example.mp3')
librosa.feature.rmse(y=y)
S,phase = librosa.magphase(librosa.stft(y))
rms = librosa.feature.rmse(S=S)
```

由于低层的特征为一维特征，即线性特征，所以这部分内容进入卷积神经网络时，需要对输入层数据进行预处理，将高度设置为 1，并且卷积和池化时其高度均须固定为 1，相当于卷积操作和池化操作过程中是在左右的方向上进行，没有上下移动。

2. MFCC 特征

梅尔频率倒频谱分析是基于人类的听觉感知设计的，人耳对低频部分的音频比较敏锐，核心思想是通过滤波器组的方式模拟人耳对不同频率音频的感知，将频谱从线性分布转换为非线性分布，具体转换方法如式(11.1)所示。梅尔频率倒谱系数(MFCC)在梅尔频率的尺度上进行频谱分析，经过倒谱分析之后，得到的结果系数即为这一帧音频的特征。一段音频的倒谱系数形成数组结果序列，通过对这些倒谱向量进行分析，就可以获得音频的质量。

$$\mathrm{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{11.1}$$

式中的 f 为输入音频实际频率，单位为 Hz。提取过程中首先对输入音频进行预加重、分帧和分窗操作，窗口大小为 512，然后对每个窗口运行 FFT 运算得到频谱，将结果取绝对值或平方值后进行 Mel 滤波运算得到 Mel 频谱，对滤波器结果取对数后进行离散余弦变换，最后取系数作为 MFCC 特征向量。经过 MFCC 计算后可以获得的音频特征为 39 个，其中包括了 12 个倒谱特征系数，12 个 Δ 倒谱特征系数，12 个 $\Delta\Delta$ 倒谱特征系数，1 个能量系数，1 个 Δ 能量系数，1 个 $\Delta\Delta$ 能量系数。

应用 librosa 库可以方便地提取上述特征，具体代码如下，其中窗口大小为 512，而特征数为 39 个。

```
y,sr = librosa.load(srcFile)
y_harmonic,y_percussive = librosa.effects.hpss(y)
tempo,beat_frames = librosa.beat.beat_track(y = y_percussive,sr = sr)
mfcc = librosa.feature.mfcc(y = y,sr = sr,hop_length = 512,n_mfcc = 39)
```

3. 心理声学特征

心理声学特征描绘的主要是人的主观感受，由于人耳的听觉特性和机制目前的研究尚未完全解释清楚，特别是受到掩蔽、非线性、双耳效应等影响，所以目前借鉴的是常用的声学模型特征，主要从响度、音调、音色的角度进行特征提取，在本系统中采用以下心理声学特征：音频响度(Loudness)代表了音频能量的强弱变化，与时域波形的振幅大小成正比；尖锐度(Sharpness)反映的是音频是否刺耳及其程度，一般以高频部分在整个音频频谱中点的比例来衡量。

使用 librosa 库提取色调质心特征(tonnetz)，代码如下，还可以对音频色度特征(chrome)进行提取，提取方法为 librosa.feature.chroma。

```
y,sr = librosa.load(srcFile)
tonnetz = librosa.feature.tonnetz(y = y,sr = sr)
```

上述 3 个特征经过处理后均保存为图片格式，处理方法是应用 python 中的 Matplotlib 库进行图片保存，其结果如图 11.2 所示。

由于保存后的图片含有坐标值和标题等文字说明，不利于卷积计算，所以在保存前将坐标禁用，同时，缩进 3 个像素进行裁剪，去除图片的边框，完整的代码如下。

```
fig,ax = plt.subplots()
plt.axis('off')
librosa.display.specshow(mfcc,sr = sr)
```

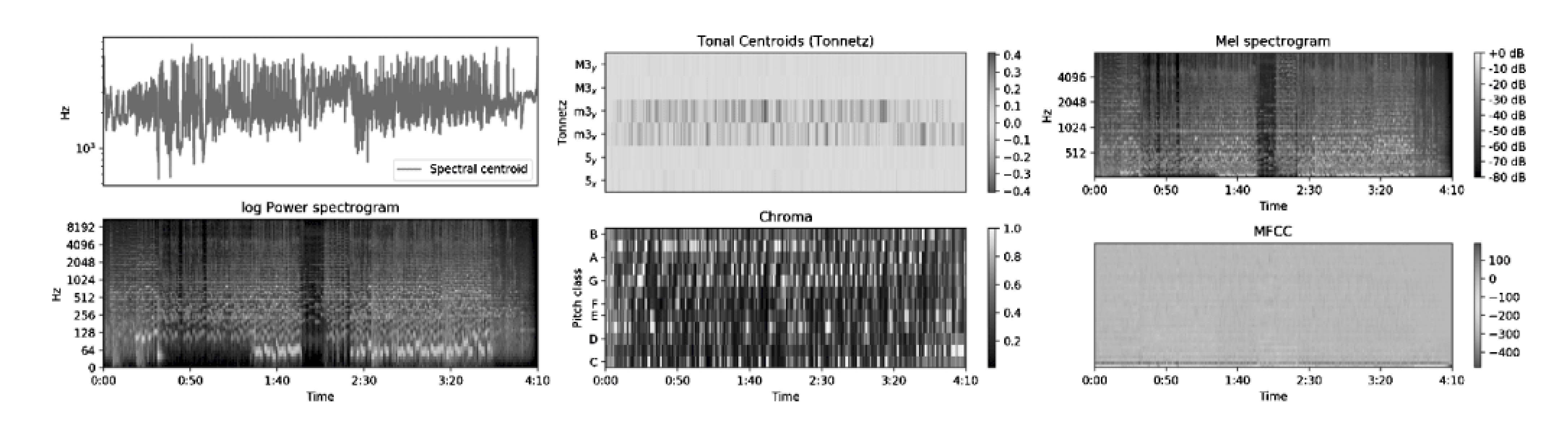


图 11.2　频谱质心、MFCC、语谱图、Mel Spectrogram 等特征

```
filename = os.path.splitext(basename(srcFile))[0]
filename = targetFolder + filename + ".png"
print('Saving output to ' + filename)
extent = ax.get_window_extent().transformed(fig.dpi_scale_trans.inverted())
plt.savefig(filename,bbox_inches = extent,pad_inches = 0)
plt.close()
img = Image.open(filename)
rect = (3,3,img.width - 3,img.height - 3)
img.crop(rect).save(filename)
img.close()
```

11.2.2　音频特征选择

为了比较相同网络结构下音频特征的区分度，为了减小网络结构带来的影响，都使用结构相同的 LeNet 网络模型，比较 Spectrogram、CQT、MFCC、色调质心特征(Tonnetz)、频谱对比度(Contrast)、音频节奏特征(Tempo)对音频质量分类的区分度。

网络的随机数种子设置为 42，训练集和测试集按照 8∶2 的比例分配样本，每批样本数为 40，训练周期(epoch)为 10，输入图像的大小处理为 200 像素×200 像素，3 通道的图片，网络中 L2 参数为 0.0001，激活函数为 RELU，学习率为 0.0001，网络初始化方法为 Xavier，采用随机梯度下降(SGD)进行优化。第一层卷积层通道为 3，输出为 50、卷积核为 5×5、步长为 1×1；第二层为最大化池化层，核大小为 2×2；第三层卷积层输出 100、卷积核为 5×5、步长为 1×1；第四层为最大化池化层，核大小为 2×2；第五层全连接层输出 500；第六层输出层损失函数为负的 Log 似然函数(negative log-likelihood)，输出个数为分类个数 3，输出层的激活函数为 Softmax。网络结构定义相关的代码如下：

```
MultiLayerConfiguration conf = new NeuralNetConfiguration.Builder()
        .seed(seed)
        .iterations(iterations)
        .regularization(true).l2(0.0001) // tried 0.0001,0.0005
        .activation(Activation.RELU)
        .learningRate(0.0001) // tried 0.00001,0.00005,0.000001
        .weightInit(WeightInit.XAVIER)
        .optimizationAlgo(OptimizationAlgorithm.STOCHASTIC_GRADIENT_DESCENT)
        .updater(Updater.NESTEROVS).momentum(0.9)
        .list()
        .layer(0,convInit("cnn1",channels,50,new int[]{5,5},new int[]{1,1},new int[]{0,0},0))
```

```
.layer(1,maxPool("maxpool1",new int[]{2,2}))
.layer(2,conv5x5("cnn2",100,new int[]{5,5},new int[]{1,1},0))
.layer(3,maxPool("maxool2",new int[]{2,2}))
.layer(4,new DenseLayer.Builder().nOut(500).build())
.layer(5,new OutputLayer.Builder(LossFunctions.LossFunction.NEGATIVELOGLIKELIHOOD)
        .nOut(numLabels)
        .activation(Activation.SOFTMAX)
        .build())
.backprop(true).pretrain(false)
.setInputType(InputType.convolutional(height,width,channels))
.build();
```

各音频特征的比较结果见表 11.1。

表 11.1 相同卷积神经网络不同音频特征表现

音频特征	准确率	精确率	召回率	F1 分值
MFCC	0.4286	0.4583	0.5000	0.4783
CQT	0.6250	0.8000	0.6667	0.7273
Spectrogram	0.6667	0.7500	0.6667	0.7059
Tonnetz	0.2857	0.2917	0.3333	0.3111
Contrast	0.7143	0.7778	0.7778	0.7778
Tempo	0.5714	0.5833	0.5556	0.5691

从表 11.1 中可以看出，语谱图特征具有较高的区分度，语谱图中信息量相对较复杂，比较适合卷积神经网络处理和分析，而 MFCC 虽然包括了 39 维的音频特征，但是其图形显示过于简单，并且各维度之间的关系相对独立，以二维的方式进行分析，提取的特征并不明显。CQT 和 Contrast 音频特征具有二维图片的关联特性，含有的信息与语谱图类似，其准确率、F1 分值等数据比语谱图略低。所以，在卷积神经网络模型的特征选择中，使用 Spectrogram、Contrast、CQT 作为模型的输入特征。

11.2.3 卷积神经网络模型训练

对音频进行特征提取并应用到卷积神经网络中，通过网络模型的机器学习，并结合不同质量音频的实际质量，对网络中的参数进行人工微调，使其在训练音频集中得到的评分结果与实际人工听评结果尽可能吻合。

基于模型的网络结构中含有 3 个相对独立的子网络，所以在模型训练过程中采用先子网络后总网络的训练过程，即先对语谱图、Tempo 等特征、低层特征图进行训练，调整其网络参数，使其结果达到最优，然后调整子网络的输出权重。下面的代码使用 UIServer 对训练过程中的模型进行可视化，并将训练后的模型保存在指定的文件。

```
StatsStorage statsStorage = new FileStatsStorage(statsFile);
int listenerFrequency = 1;
network.setListeners(new StatsListener(statsStorage,listenerFrequency));
UIServer uiServer = UIServer.getInstance();
uiServer.attach(statsStorage);
```

训练过程启动后，可以在浏览器中输入 http://localhost:9000/查看模型结构及各项参数指标的图形显示。图 11.3 是网络模型在训练过程中对音频质量进行评分的结果和迭代过程，如果模型中参数设置失效或不合理，将容易产生梯度消失或学习率过低等问题，从而导致最终的分类结果准确度很差。由于每次训练的时间大约在 1h，通过可视化的界面可以提前发现梯度消失等问题，直接中止训练过程，重新进行参数调整，这样可提高模型训练的效率。

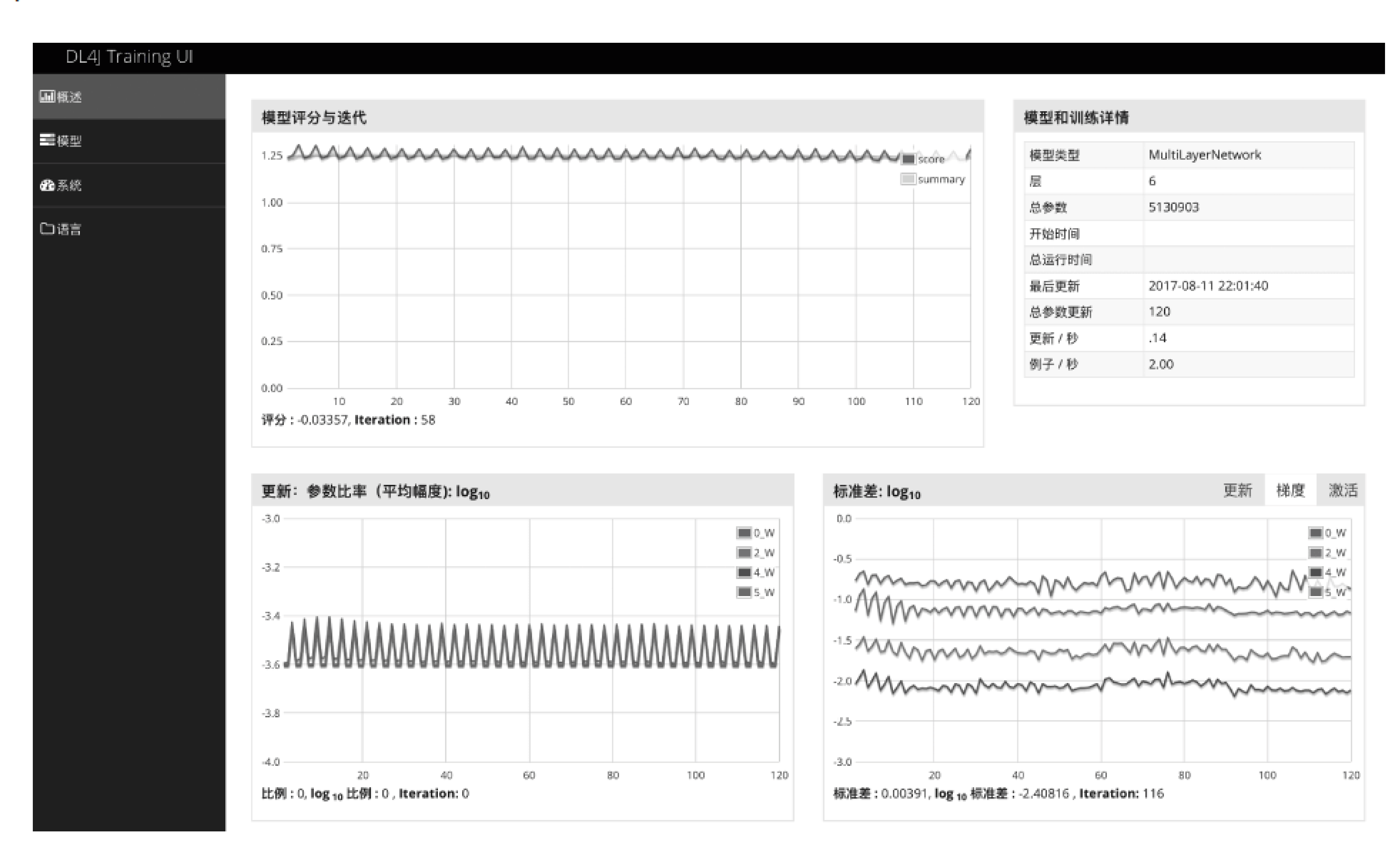


图 11.3　卷积神经网络模型评分与迭代过程

图 11.4 是语谱图子网络在训练过程中的各权重的更新值与参数值之比。从图 11.4 中可以看出，随着迭代次数的增加，网络中的权重更新值与参数的比值变化较为剧烈，特别是 5W 对应的第 5 层(全连接层)的比率变化曲线，表示模型结构不稳定，学习不到有用特征。

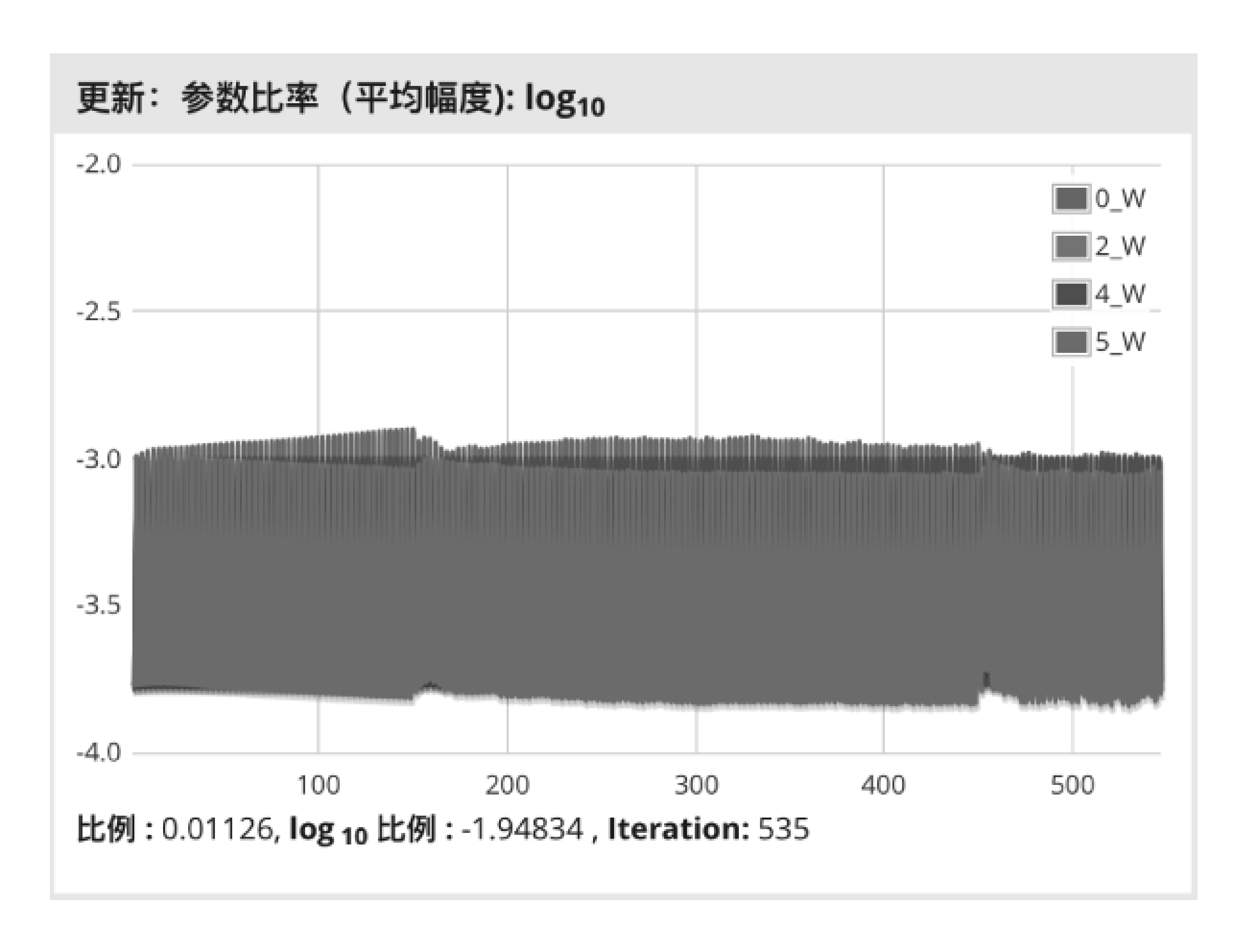


图 11.4　语谱图子网络在训练过程中的各权重的更新值与参数值之比

在 Web 可视化界面的导航中单击“模型”，进入模型结构和参数查看页面，如图 11.5 所示，左上侧是模型的结构，单击某个节点将显示此层的详情和对应各参数的可视化图表结果。图 11.5 中为卷积层 1 对应的信息，可以看到其内核大小为 5 像素×5 像素，步长为 1 像素×1 像素，无填充，激活函数为 relu，还可以看到参数更新的比率幅度和激活函数结果变化情况，最右下方是学习率的变化情况，模型中此参数为固定值，所以其值为直线，在其他应用中可以指定学习率为动态值。

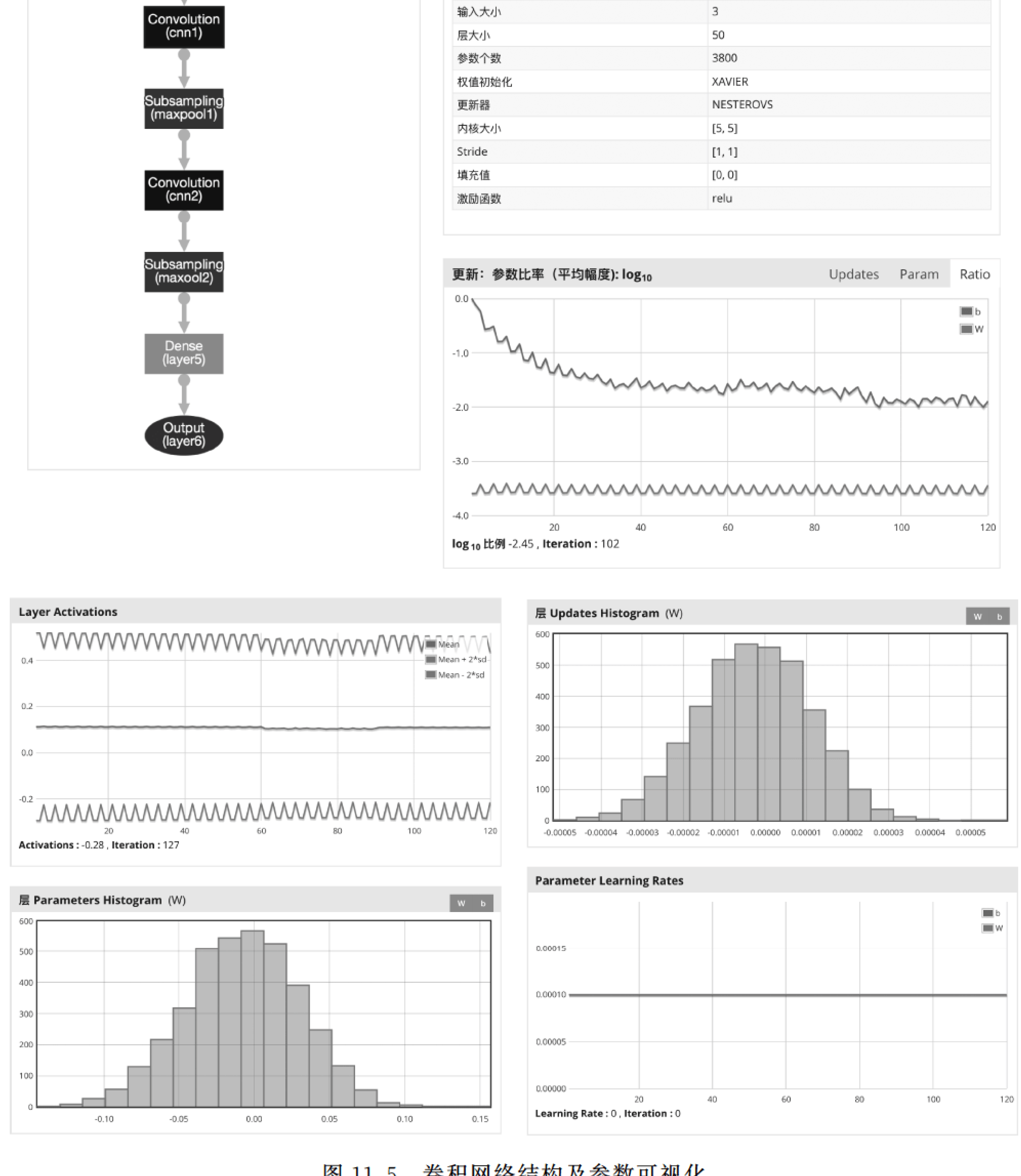


图 11.5　卷积网络结构及参数可视化

11.2.4 模型参数调优

卷积神经网络模型中的主要可调参数为：输入层特征图片处理的长度（Height）和宽度（Width）、输入层图片的通道数（Channel）、动量（Momentum）参数、迭代次数（Iterations）、正则化参数 L2、学习率（Learning Rate）、卷积层卷积核（Kernel）大小、卷积层输入输出、步长（Stride）、填充（Padding）、偏移（Bias）、池化层大小（Kernel）、全连接层输出数量、激活函数（Activation）。另外，针对不同的网络结构进行调整，如层数进行增删等，对模型进行性能调优。

音频评价卷积模型的训练过程采用随机梯度下降（Stochastic Gradient Descent，SGD）的方式进行迭代参数更新，迭代周期次数为 50 次。每个迭代周期是指在模型训练过程中完整地遍历一次训练集，从而使模型调整参数。首先从训练音频集中提取所有音频，对音频进行特征提取操作，提取语谱图特征、提取 MFCC 特征并存储为 png 图片格式，另将低层特征存储为 csv 文本格式。

输入层的特征图片尺寸为 100 像素×100 像素时，比 200 像素×200 像素的训练时间相对较短，但是准确率下降，而长宽的像素越大，需要的硬件配置也越高，特征图片大小超过 250 像素×250 像素之后，模型的参数变多，模型文件变大，对准确率改进有限。

通过对输入的特征图片进行转换增加训练样本数量，使用翻转变换、随机翻转变换、扭曲变换、颜色转换变换对特征图进行处理，每执行一次变换，重新训练 50 个迭代周期次数。经过实验对比，发现增加颜色变换操作后，模型的训练时间增长，而分类准确率和精确率反而下降，说明在训练过程中并不需要颜色转换变换，刻意增加参数不一定效果更好。

基于 LeNet 网络对语谱图进行训练，在原结构的全连接层之前增加两个层：卷积层（输出个数为 500，步长为 5 像素×5 像素，填充 1 像素×1 像素）和池化层（核大小为 2 像素×2 像素），模型的效果不但没有提升，反而有所下降。说明单纯增加层数对模型优化没有太多帮助。

激活函数随时间变化的曲线可用于检验激活函数的消失或膨胀。理想情况下，曲线应该随着时间变化越来越稳定，其标准差的取值范围为 0.5～2.0，如果严重超过这个范围，说明出现了权重值初始化不合理、正则化过度或数据标准化不足等问题，也有可能是学习率设置不当。

从模型训练过程的可视化可以看出模型中更新器的更新模式选择是否合理，或者模型参数设置是否合适。更新值与参数的比例图可用于设置学习率，一般的比例应该在 0.001，即在图中的坐标系下应该在－3.0 附近，如果此比例在训练迭代过程中出现了大幅上升，说明发生了梯度膨胀。

将图片等特征作为输入层信号传入卷积神经网络模型中，并对模型进行参数修改，对训练后的模型使用验证集进行验证，对模型的性能进行验证，并输出实际结果和预测结果的对比。模型的性能指标包括模型准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 分值（F1 Score）。如果分值与实际人工听评结果值相差过大或模型的性能指标过低，则执行惩罚，以修正网络参数，重复多次，直到分值与实际人工听评值接近一致或模型性能指标超过 80％。

经过对模型进行参数调整和迭代训练，并在训练过程中对卷积神经网的层数和卷积核

大小进行调整，语谱图子网络和 Tempo 子网络的层数为 7 层，低层特征子网络的层数为 5 层，训练总参数最高达到 5 131 404 个，其中语谱图的区分度最高，得到准确率为 87.5%、精确率为 91.67%、F1 分值为 89.53%的卷积神经网络模型子网络模型。在调整参数过程中发现，在训练样本不变的情况下随着模型层数的增加，训练时间大幅增加，但是模型的精度并不升高，反而有所下降。

11.3　性能验证

将卷积神经网络模型算法与不同的分类算法进行效果比较，由于线性支持向量机分类算法 LinearSVC 是线性 SVM 算法的一种，并且在分类效果上较优秀，而 k 近邻(KNN)在机器学习算法中也具有简单高效的特点，所以应用 LeNet 网络的卷积神经网络模型算法与 KNN 算法、LinearSVC 进行分类效果对比，代码如下：

```
dataImages = np.array(inputImages)
labels = np.array(inputLabels)
(trainImgs,testImgs,trainLabel,testLabel) = train_test_split( dataImages,labels,test_size =
0.20,random_state = 42)
print("Training model...")
modelKNN = KNeighborsClassifier(n_neighbors = args["neighbors"],n_jobs = 4)
modelKNN.fit(trainImgs,trainLabel)
acc = modelKNN.score(testImgs,testLabel)
print("[INFO] k-NN model accuracy: {:.4f}".format(acc))
labelList = LabelEncoder()
labels = labelList.fit_transform(labels)
print("[INFO] Evaluating k-NN model...")
predictions = model.predict(testImgs)
print(classification_report(testLabel,predictions,target_names = labelList.classes_))
modelSVC = LinearSVC()
modelSVC.fit(trainImgs,trainLabel)
acc = modelSVC.score(testImgs,testLabel)
print("[INFO] linearSVC accuracy: {:.4f} % ".format(acc))
print("[INFO] Evaluating linearSVC model...")
predictions = modelSVC.predict(testImgs)
print(classification_report(testLabel,predictions,target_names = labelList.classes_))
```

在对比实验过程中，分别设置 KNN 算法中的 k 值对其进行优化，选取其中使 KNN 算法取得较高评估效果的 k 值作为最终参数值，并记录算法的分类表现。KNN 和 LinearSVC 算法由算法库 sklearn 提供，集成于 OpenCV-Python 库中，供 Python 代码调用。对相同音频训练集和测试库进行特征提取后，将其应用于不同的算法中进行分训练和评估，代码如下：

```
le = LabelEncoder()
labels = le.fit_transform(labels)
print("[INFO] constructing training/testing split...")
(trainData,testData,trainLabels,testLabels) = train_test_split(
    np.array(data),labels,test_size = 0.25,random_state = 42)
print("[INFO] training Linear SVM classifier...")
model = LinearSVC()
```

```
model.fit(trainData,trainLabels)
print("[INFO] evaluating classifier...")
predictions = model.predict(testData)
print(classification_report(testLabels,predictions,target_names = le.classes_))
```

将得到的分类评估结果与卷积神经网络模型的结果进行对比，如图 11.6 所示。由于卷积神经网络评分模型中主要采用 Spectrogram 和 CQT 音频特征作为模型输入，所以本实验中也采用上述两种音频特征的图片列表作为各算法的输入信号。

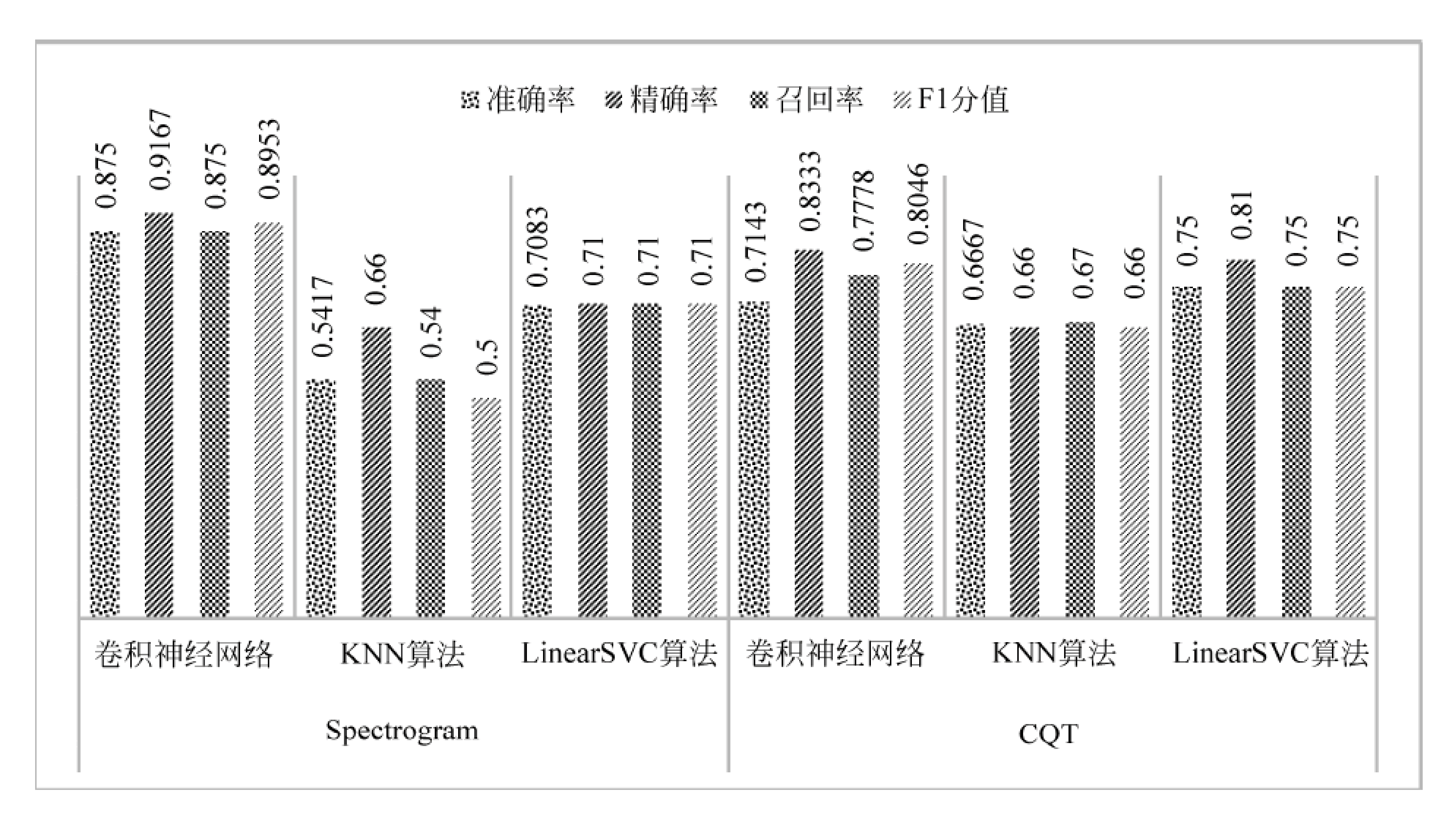


图 11.6　卷积神经网络与其他分类算法效果比较

从结果可以看到，卷积神经网络算法在 Spectrogram 和 CQT 特征上的表现要优于 KNN 和 LinearSVC 算法，但结果并不明显，这可能与音频的样本数较少有关，而卷积神经网络与其他分类算法相比，对异常特征值的过滤能力较差，在小样本的测试中并不能展现其深度学习的能力，通过在实际系统中增加音频训练的样本数量和动态调用音频训练库，可以将其与其他分类算法的差距进一步拉大，表现也将更加优秀。

模型采用了音频低层特征、Tempo 等音频特征、音频语谱图作为卷积神经网络的输入特征，采用机器学习的方式从音频特征中抽取特征参数，经过反复训练和调优，逐步与实际的人工听评结果相一致。由于采用的音频特征数量较多，依据传统的神经网络原理，其运算量较大，而卷积神经网络具有局部感知、共享权值等优点，用其代替传统的神经网络，不仅运算量骤减，而且音频特征并未消失，不影响特征提取和音频评价结果的准确性。

局限于训练音频库的训练音频数量，某些特征并不明显，会导致模型训练无法达到 90%以上的分类精度，为了弥补这一缺陷，模型中参数的调整除在模型训练过程中进行，在模型应用中也对其进行修正，如果发现有评分结果失误的音频，则将其重新标记评分等级，并提交至训练音频库中，定期运行模型训练程序对模型进行学习训练。这样将使模型的精度和分类区分度随着训练音频库的增长而不断提高。

参考文献

[1] Ramesh Sharda, Dursun Delen, Efraim Turban. 商务智能：数据分析的管理视角[M]. 3 版. 赵卫东，译. 北京：机械工业出版社，2014.

[2] 赵卫东. 商务智能[M]. 4 版. 北京：清华大学出版社，2016.

[3] 卢辉. 数据挖掘与数据化运营实战：思路、方法、技巧与应用[M]. 北京：机械工业出版社，2013.

[4] 赵卫东，赵洪博. 基于项目沉浸式的数据分析类课程教学研究[J]. 计算机教育，2017，270：58-61.

[5] [美]纳撒尼尔·林. 大数据商业分析：整合大数据与业务流程的高级商业分析指南[M]. 北京：人民邮电出版社，2016.

[6] Danie T Larose, Chantal D Larose. 数据挖掘与预测分析[M]. 2 版. 王念滨，宋敏，裴大茗，译. 北京：清华大学出版社，2017.

[7] 张良均，陈俊德，刘名军，等. 数据挖掘实用案例分析[M]. 北京：机械工业出版社，2016.

[8] James Evans. Business Analytics[M]. New York：Pearson Education Limited，2017.

[9] 陈春宝，阙子扬，钟飞. 大数据与机器学习实践方法与行业案例[M]. 北京：机械工业出版社，2017.

[10] 周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[11] Jeffrey D Camm, James J Cochran, Michael J Fry, et al. Essentials of Business Analytics[M]. Boston：Cengage Learning，2015.

[12] 吴岸城. 神经网络与深度学习[M]. 北京：电子工业出版社，2016.

[13] Jared Dean. Big data, data mining and machine learning：value creation for business and machine learning[M]. Hoboken：John Wiley & Sons Ltd.，2014.

[14] 郑泽宇，顾思宇. TensorFlow 实战 Google 深度学习框架[M]. 北京：电子工业出版社，2017.

[15] 焦李成. 深度学习优化与识别[M]. 北京：清华大学出版社，2017.

[16] 王琛，胡振邦，高杰. 深度学习原理与 TensorFlow 实践[M]. 北京：电子工业出版社，2017.

[17] 张文彤，钟云飞. IBM SPSS 数据分析与挖掘实战案例精粹[M]. 北京：清华大学出版社，2013.

[18] 薛薇. 基于 SPSS 的数据分析[M]. 3 版. 北京：中国人民大学出版社，2014.

图书资源支持

感谢您一直以来对清华版图书的支持和爱护。为了配合本书的使用，本书提供配套的资源，有需求的读者请扫描下方的“书圈”微信公众号二维码，在图书专区下载，也可以拨打电话或发送电子邮件咨询。

如果您在使用本书的过程中遇到了什么问题，或者有相关图书出版计划，也请您发邮件告诉我们，以便我们更好地为您服务。

我们的联系方式：

地　　址：北京海淀区双清路学研大厦 A 座 707

邮　　编：100084

电　　话：010－62770175－4604

资源下载：http://www.tup.com.cn

电子邮件：weijj@tup.tsinghua.edu.cn

QQ：883604（请写明您的单位和姓名）



书圈

用微信扫一扫右边的二维码，即可关注清华大学出版社公众号“书圈”。